Three families of Asgard archaeal viruses identified in metagenome-assembled genomes

Sofia Medvedeva, Jiarui Sun, Natalya Yutin, Eugene V. Koonin, Takuro Nunoura, Christian Rinke, Mart Krupovic

Three families of Asgard archaeal viruses identified in metagenome-assembled genomes

Sofia Medvedeva[1,2#], Jiarui Sun[3], Natalya Yutin[4], Eugene V. Koonin[4], Takuro Nunoura[5]*, Christian Rinke[3]*, Mart Krupovic[1]*

[1] Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris, France
[2] Center of Life Science, Skolkovo Institute of Science and Technology, Moscow, Russia
[3] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD, Australia
[4] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[5] Research Center for Bioscience and Nanoscience (CeBN), Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan

* Correspondence to
Takuro Nunoura, E-mail: takuron@jamstec.go.jp
Christian Rinke, E-mail: c.rinke@uq.edu.au
Mart Krupovic, E-mail: mart.krupovic@pasteur.fr

Present address:
Institut Pasteur, Université Paris Cité, CNRS UMR6047, Evolutionary Biology of the Microbial Cell Unit, Paris, France

Abstract

Asgardarchaeota harbour many eukaryotic signature proteins and are widely considered to represent the closest archaeal relatives of eukaryotes. Whether similarities between Asgard archaea and eukaryotes extend to their viromes remains unknown. Here we present 20 metagenome-assembled genomes of Asgardarchaeota from deep-sea sediments of the basin off the Shimokita Peninsula, Japan. By combining a CRISPR spacer search of metagenomic sequences with phylogenomic analysis, we identify three family-level groups of viruses associated with Asgard archaea. The first group, verdandiviruses, includes tailed viruses of the class *Caudoviricetes* (realm *Duplodnaviria*); the second, skuldviruses, consists of viruses with predicted icosahedral capsids of the realm *Varidnaviria*; and the third group, wyrdviruses, is related to spindle-shaped viruses previously identified in other archaea. More than 90% of the proteins encoded by these viruses of Asgard archaea show no sequence similarity to proteins encoded by other known viruses. Nevertheless, all three proposed families consist of viruses typical of prokaryotes, providing no indication of specific evolutionary relationships between viruses infecting Asgard archaea and eukaryotes. Verdandiviruses and skuldviruses are likely to be lytic, whereas wyrdviruses potentially establish chronic infection and are released without host cell lysis. All three groups of viruses are predicted to play important roles in controlling Asgard archaea populations in deep-sea ecosystems.

Introduction

Asgard archaea are an expansive group of metabolically versatile archaea that thrive primarily in anoxic sediments around the globe[1-9]. Based on phylogenomic analyses, asgardarchaea have been originally classified into multiple phylum-level lineages, including Lokiarchaeota, Thorarchaeota, Odinarchaeota, Heimdallarchaeota, Helarchaeota, Sifarchaeota, Wukongarchaeota and several others, most of which were named after Norse gods[1,5,7,9-12]. Recently, taxonomic rank normalization using relative evolutionary divergence has suggested that asgardarchaea represent a phylum, tentatively named Asgardarchaeota, including the classes Lokiarchaeia, Thorarchaeia, Odinarchaeia, Heimdallarchaeia, Sifarchaeia, Hermodarchaeia, Sifarchaeia, Baldrarchaeia, Wukongarchaeia and Jordarchaeia, with the other lineages classified as lower-rank taxa within the classes[8,13]. The vast majority of asgardarchaea have been discovered through metagenomics, whereas only one species has been isolated and successfully grown in the laboratory[14]. Asgardarchaea gained prominence due to their inferred key role in the origin of

eukaryotes[15]. Indeed, Heimdallarchaeia form a sister group to eukaryotes in most phylogenetic analyses[1,5], although alternative phylogenies have also been presented [16,17]. Compared to other archaea, asgardarchaea encode a substantially expanded set of eukaryotic signature proteins, including many proteins implicated in membrane trafficking, vesicle formation and transport, cytoskeleton formation, the ubiquitin network and other processes characteristic of eukaryotes[1,5]. A tantalizing question is whether eukaryotes also inherited viruses and other types of mobile genetic elements (MGE) from the asgardarchaea?

Viruses infecting archaea are remarkably diverse, both in terms of their genome sequences and virion structures[18-20]. Some archaeal viruses, in particular those with icosahedral virions, are evolutionarily related to bacterial and eukaryotic viruses, but the majority of archaeal virus groups are specific to archaea, with no identifiable relatives in the two other domains. Archaea-specific viruses often have odd-shaped virions, which resemble lemons, champagne bottles or droplets[19]. Most archaeal viruses have been thus far isolated from hyperthermophilic or halophilic hosts, with only a handful of virus species described for methanogenic and ammonia-oxidizing mesophilic archaea[18]. No viruses infecting asgardarchaea have been isolated, primarily due to the inherent difficulty to propagate asgardarchaeal hosts. Nevertheless, analysis of the CRISPR-Cas loci in the genomes of asgardarchaea revealed a remarkable diversity of defense systems in these organisms[21], implying a rich asgardarchaeal virome. CRISPR arrays are archives of past viral/MGE encounters, which can be harnessed to uncover the associations between viruses and their hosts. Indeed, matching the CRISPR spacers from a known organism to viruses with unknown hosts is widely used for host assignment for viruses discovered by metagenomics and arguably is the most straightforward and efficient approach to identify the hosts of viruses infecting prokaryotes[22].

Here we harness CRISPR spacer sequences from the sequenced asgardarchaeal genomes to search for viruses infecting these organisms and describe three distinct family-level groups of Asgard-associated viruses, all of which display typical features of viruses infecting bacteria or archaea.

Results

To search for viruses infecting asgardarchaea, we sequenced 12 metagenomes constructed from total environmental DNA directly extracted from the subseafloor sediments originating off the Shimokita Peninsula, Japan (site C9001, water depth 1,180 m)[23] and representing different sediment depths ranging from 0.91 to 363.3 meters below the seafloor (mbsf). Asgardarchaeal metagenome-assembled genomes (MAGs) were assembled from 7 of the 12 samples. A total of 20 Asgardarchaeota MAGs, including 12 Lokiarchaeia, 2 Thorarchaeia and 6 Heimdallarchaeia were analyzed in this study (Fig. 1), with an estimated completeness of 65.95±17.93, estimated quality of 54.68±15.78, and GC content of 32.78±3.99 (Table S1).

To assign the putative viral genomes to asgardarchaeal hosts, we assembled a dataset of CRISPR spacers from asgardarchaeal MAGs assembled in this study as well as those reported previously (Fig. 2, Table S2). Analysis of the Asgard CRISPR arrays allowed us to define Asgard-specific CRISPR repeat sequences, which were used to identify additional CRISPR arrays from the contigs obtained from our sediment samples (Fig. 2a; see Methods for details). In total, our CRISPR spacer dataset (Fig. 2b, Supplementary Data 1), included 2,532 spacers assigned to different Asgardarchaota lineages, with Lokiarchaeia contributing the highest number of spacers (Fig. 2c). All spacers were then used to search for protospacers in the 20 assembled MAGs and putative virus genomes from our dataset as well as contigs from GenBank and JGI sequence databases. In total, 14 contigs could be assigned to asgardarchaeal hosts based on CRISPR spacer matches with estimated false positive rate of 0.00276 (see Methods for details, Extended Data 1). By contrast, none of the contigs was targeted by spacers from the CRISPRHost database (675,911 spacers) and only two, likely false positive, spacers were identified in the CRISPR Spacer Database and Exploration Tool (11,674,395 spacers; see Methods).

Eight of the putative viral genomes originated from the metagenomes sequenced in this study (from depths ranging from 0.91 to 87.7 mbsf), whereas six additional genomes were recovered from datasets sequenced previously, including anoxic subseafloor sediment samples from two Pacific Ocean sites, the Hikurangi Subduction Margin (144.3 mbsf)[8] and Cascadia Margin (2 mbsf)[24], and one Indian Ocean site, Sumatra Forearc

(1.6-6.07 mbsf; PRJNA367446). Each genome was targeted (≥90% identity between spacer and protospacer) by 1-6 CRISPR spacers assigned to putative Asgard classes Lokiarchaeia, Thorarchaeia, and Heimdallarchaeia (Fig. 1, Tables S3 and S4). Based on the conservation of viral hallmark proteins, including major capsid proteins (MCP), seven of the genomes were unequivocally identified as belonging to viruses from three unrelated groups (Fig. 3-5), whereas the remaining seven could represent either unknown viruses or other types of mobile genetic elements (MGEs; Fig. S1, see below). The viral dataset was further enriched by searching the JGI and GenBank sequence databases for related contigs. As a result, we collected 21 contig representing three groups of asgardarchaeal viruses originating from seven geographically remote locations (Fig. 6a). Further analyses using CRISPR-based (SpacePHARER) and kmer-based (WiSH and PHIST) methods confirmed the association of the 21 contigs with asgardarchaeal hosts (see Methods, Tables S3 and S5). Network analysis of these viral genomes together with the bacterial and archaeal virus genomes available in RefSeq showed that the three Asgard virus groups are disconnected from each other as well as from other known viruses (Fig. 6b).

Verdandiviruses are tailed viruses of the *Caudoviricetes*

Three of the genomes, VerdaV1, -2 and -3, assembled from the Shimokita dataset, encode the hallmark proteins specific to viruses of the class *Caudoviricetes*, the most widespread, environmentally abundant and genetically diverse group of viruses[25,26]. Members of the *Caudoviricetes* infect bacteria and archaea, and have characteristic virions consisting of an icosahedral capsid and a helical tail attached to one of the capsid vertices. *Caudoviricetes* together with eukaryotic herpesviruses form the realm *Duplodnaviria*[27]. Similar to previously characterized bacterial and archaeal *Caudoviricetes*[18], each of the three virus genomes encodes the HK97-like MCP, large subunit of the terminase (genome packaging ATPase-nuclease), the portal protein as well as several other structural proteins, including tail components (Fig. 3a). Structural modeling of the VerdaV1 MCP using RoseTTAFold[28], yielded a model with the canonical HK97-like fold (Fig. 3b).

The VerdaV1 (19.9 kb) and VerdaV2 (19.5 kb) genomes were assembled as circular contigs (Table S3), suggesting that they correspond to complete, terminally redundant virus genomes, whereas VerdaV3 (15.4 kb) likely represents a partial genome. Comparative genomics analysis showed that VerdaV1, -2 and -3 belong to the same virus group (Fig. 3a), which we refer to as 'verdandiviruses' (for Verdandi, one of the three Norns, the most powerful beings in Norse mythology that govern the lives of gods and mortals). Given that the three genomes were identified in different sediment depths of offshore Shimokita Peninsula (at 18.7, 59.5 and 87.7 mbsf; Table S3), we addressed the possibility that related viruses could also be detected in samples from other depths. To this end, we performed BLASTP searches (E-value cutoff of 1e-05) queried with the corresponding MCP sequences against the assembled sequences from samples retrieved along the depth gradient. The analysis yielded six additional viral contigs and showed that related MCPs (and viruses) are also present in sediment samples from 0.9, 9.3 and 30.8 mbsf, indicating a broad distribution of verdandiviruses through the sediment column. One of the retrieved contigs (VerdaV4; 20.6 kb) was found to be circular and thus also represents a complete virus genome. Furthermore, searches against the GenBank and JGI sequence databases yielded eight hits to virus-like contigs affiliated to asgardarchaea (Table S3). Five of the MCPs were encoded within large contigs (>10 kb) including two (Ga0114925_10000341 and Ga0114923_10001063 [VerdaV5 and VerdaV6, respectively]) circular ones. Notably, VerdaV5 was also targeted by a CRISPR spacer from a recently described Asgardarchaeota MAG[1], further supporting the host assignment for the Verdandivirus group (Fig. 3a, Table S3). Finally, we identified a partial provirus related to verdandiviruses integrated in a large genomic contig of Lokiarchaeia (Extended Data 3). The same contig contained cellular genes unequivocally belonging to Asgardarchaeota (98-99% protein identity), including those encoding ribosomal proteins. Verdandivirus contigs were recovered from four geographically remote sampling sites (Fig. 6a) and maximum likelihood phylogenetic analysis of the MCPs revealed an overall biogeographic clustering of verdandiviruses (Fig. 3c).

Although virus-encoded homologs could be identified for only 2% (7 out of 298) of verdandivirus proteins by blastp searches (E=1e-05), functional annotation for some of the other proteins was enabled by sensitive profile-profile comparisons, in which HK97-like MCP and TerL MCP were readily identified with highly

significant scores (Fig. S2, Table S5). Verdandiviruses display considerable sequence conservation within the capsid formation and genome packaging modules (Fig. 3a). Notably, upstream of the MCP, all viruses carry a gene encoding a putative capsid stabilization protein distantly related to the corresponding protein of marine siphoviruses, e.g., TW1 [29], which might be important for maintaining capsid integrity under high hydrostatic pressures of the deep-sea ecosystems. Verdandiviruses encode tail proteins most closely similar to those of siphoviruses, including the major tail protein, tail tape measure protein and a baseplate hub protein with an adhesin domain, suggesting that verdandiviruses possess flexible, noncontractile tails. In most of these viruses, the tape measure protein is relatively short (median length 259 aa), suggesting that the tails themselves are short as well. Assuming ~1.5 Å of tail length per amino acid residue of tape measure protein[30,31], the median tail length of verdandiviruses is predicted to be ~39 nm. The genes encoding tail proteins display low sequence conservation, whereas gene contents downstream of the tail modules are distinct in most of the identified viruses (Fig. 3a). Differences in the tail modules might correspond to different host ranges of the corresponding viruses. Indeed, whereas VerdaV1 and VerdaV5 were matched by CRISPR spacers from Lokiarchaeia, VerdaV2 and VerdaV3 are predicted to infect Thorarchaeia (Table S3). The distinct conservation patterns of the capsid and tail modules likely reflect modular evolution of these asgardarchaeal virus genomes, a common trait in bacterial and archaeal members of the *Caudoviricetes*[32,33].

Verdandiviruses do not encode any proteins implicated in virus genome replication, and thus, can be predicted to fully rely on the hosts for this process. Complete dependence on the host replication machinery is common among archaeal viruses with small and mid-sized genomes[34], in particular, for the tailed viruses of haloarchaea and methanogens in the families *Saparoviridae*, *Suolaviridae*, *Leisingerviridae* and *Anaerodiviridae*[35]. Verdandiviruses encode predicted DNA-binding proteins with Zn-finger and helix-turn-helix motifs, which could participate in recruitment of the host replication and transcription machineries. Indeed, in the case of hyperthermophilic archaeal virus SIRV2, a viral helix-turn-helix protein has been shown to recruit the host DNA sliding clamp protein, a known interaction partner for many other components of the host replisome[36].

All complete verdandivirus genomes encompass arrays of short genes encoding predicted small, poorly conserved proteins, some of which are adjacent to genes encoding predicted transcriptional regulators (Fig. 3a). Similar to many bacterial viruses of the *Caudoviricetes*[37,38], some of these small, fast evolving genes could encode antidefense, in particular, anti-CRISPR proteins.

Skuldviruses: tailless icosahedral viruses of asgardarchaea

One of the contigs from asgardarchaeal MAGs assembled from anoxic sediments from the Hikurangi subduction margin[8], SkuldV1, targeted by two asgardarchaeal spacers (Table S3), encodes a double jelly-roll (DJR) MCP (Fig. 4a), a hallmark of viruses of the realm *Varidnaviria*, an expansive assemblage of viruses evolutionarily and structurally unrelated to duplodnaviruses[27]. Varidnaviruses are environmentally abundant and infect hosts from all domains of life [27,39]. Sequence searches with the SkuldV1 MCP led to the identification of two additional contigs, one from our dataset (SkuldV2) and the other one from the GenBank database (JABUBK010000319, hereinafter referred to as SkuldV3). The latter contig was obtained from subseafloor sediments from the Cascadia Margin (Ocean Drilling Program, site 1244) in the Pacific Ocean[24] and is targeted by four spacers from our dataset. SkuldV1 and SkuldV3 were assembled as circular contigs and thus correspond to complete virus genomes (Fig. 4a). In addition to the DJR MCP, the vast majority of varidnaviruses encode genome packaging ATPases of the FtsK-HerA superfamily[40]. A homolog of such ATPase was identified in all three SkuldV genomes, indicating that they are bona fide members of the realm *Varidnaviria*. We propose referring to this group of viruses as 'Skuldviruses' (for Skuld, another of the Norns).

The vast majority of skuldvirus proteins (97%; 66 of the 68 proteins) show no similarity to proteins encoded by other known viruses. Network analysis using CLANS[41] showed that the skuldvirus MCPs form a cluster separate from the previously characterized[42] groups of DJR MCPs (Fig. 4b). Nevertheless, profile-profile comparisons showed that they are most closely related to the corresponding proteins of prokaryotic viruses of the families *Corticoviridae* (bacteriophage PM2: HHsearch probability of 98.3), *Turriviridae* (archaeal virus STIV: HHsearch probability of 97.8)

and *Tectiviridae* (bacteriophage PRD1: HHsearch probability of 96.2) (Fig. S3), whereas eukaryotic viruses with DJR MCPs were recovered with considerably lower scores (*Phycodnaviridae*, Pyramimonas orientalis virus: HHsearch probability of 83.4). Structural comparison of the SkuldV1 MCP model obtained using RoseTTAFold[28] (Fig. 4b) further showed that it is most similar to the MCP of Pseudoalteromonas phage PM2 [43], a prototype of the *Corticoviridae*[44], which is widespread in marine ecosystems[45]. Nevertheless, skuldviruses do not share genes with other known viruses other than those encoding the MCP and the genome packaging ATPase. Corticoviruses, turriviruses and tectiviruses belong to the class *Tectiliviricetes*[27]. We propose that skuldviruses represent a separate virus family within the *Tectiliviricetes*.

Corticoviruses employ a rolling circle mechanism for genome replication and encode characteristic HUH superfamily endonucleases[46]. No such genes or other putative replication genes were identified in skuldviruses. Instead, all three skuldviruses encode a protein related to the A subunit of type IIB topoisomerases, such as topoisomerase VI. The latter enzyme consists of two distinct subunits, A and B, with the catalytic tyrosine residue responsible for DNA nicking located in the A subunit. Standalone A subunits, dubbed Topo mini-A, have been recently discovered in diverse bacterial and archaeal MGEs[47], but the functions of these proteins remain unknown. In the maximum likelihood phylogeny of Topo mini-A homologs, skuldvirus proteins form a clade with homologs from methanogenic and ammonia-oxidizing archaea (Extended Data 4). The skuldviral Topo mini-A might function as a replication protein, possibly initiating the rolling circle replication of the circular skuldvirus genomes, in a manner analogous to the HUH endonuclease.

Wyrdviruses are related to spindle-shaped archaeal viruses

One of the contigs from asgardarchaeal MAGs assembled from the Hikurangi Subduction Margin[8], WyrdV1 (15,570 bp), targeted by one CRISPR spacer from our collection, was found to encode two homologs of the MCPs specific to spindle-shaped archaeal viruses[48]. In particular, the closest homolog was found in haloarchaeal virus His1 (family *Halspiviridae*; Extended Data 5)[49]. His1-like MCPs are ~80 aa-long and contain two hydrophobic segments predicted to be membrane spanning domains[48] and in all spindle-shaped viruses playing key roles in virion structure and assembly[50]. BLASTP searches using the WyrdV1 MCP as a query against the JGI and GenBank databases identified eight additional contigs (Fig. 5). All these contigs shared several genes encoding a morphogenetic module, including the MCP and receptor-binding adhesin, which in fuselloviruses and halspiviruses is located at one of the pointed ends of the virion[51,52]. In addition, all these viruses encode a AAA+ ATPase, which in profile-profile comparisons showed the highest similarity to the morphogenesis (pI) proteins of bacterial filamentous phages (order *Tubulavirales*)[53]. This ATPase is responsible for the extrusion of the viral genome through the cellular membrane during virion assembly[54]. We propose referring to this group of viruses as 'wyrdviruses' (for Wyrd [Urðr], the third Norn). The homology between the tubulaviral and wyrdviral ATPases suggests that virion assembly of wyrdviruses is mechanistically similar to the extrusion of filamentous bacteriophages. Indeed, spindle-shaped viruses infecting other archaea are released from the host without causing the cell lysis through a budding-like mechanism[50,55,56]. Notably, some spindle-shaped viruses encode a single MCP (e.g., halspiviruses), whereas others encode two paralogous MCP (e.g., fuselloviruses). Similarly, wyrdviruses encode either one or two MCP paralogs (Fig. 5).

Similar to halspiviriruses and thaspiviruses, two of the contigs, Ga0209633_10003833 and Ga0209976_10001089, to which we refer to as WyrdV2 and WyrdV3, respectively, contained terminal inverted repeats (TIR), indicating that these are (nearly) complete genomes. By contrast, contigs JABUBK010000290 and JABUBK010000290, herein referred to as WyrdV4 and WyrdV5, respectively, contained direct terminal repeats, indicative of the completeness and circular structure of the genomes, resembling fuselloviruses. Consistent with the different genome structures, WyrdV2 and WyrdV3 (and halspiviruses) encode protein-primed family B DNA polymerases, whereas WyrdV4 and WyrdV5 encode rolling circle replication initiation endonucleases of the HUH superfamily (Fig. 5). Notably, WyrdV4, in addition, encodes Topo mini-A. The latter does not cluster with homologs from skuldviruses, but instead forms a clade with the Topo mini-A from the unassigned asgardarchaeal MGEs 10H_0 and 7H_42 (Extended Data 4). Contigs Ga0209976_10001631

(WyrdV6) and Ga0209976_10001236 (WyrdV7) also encode DNA polymerases and likely are linear, nearly complete viral genomes (Fig. 5). Remarkably, the DNA polymerase encoded by WyrdV7 is not orthologous to those of WyrdV2, WyrdV3 and WyrdV6 (Fig. 5). The latter group is most closely related (~30% identity) to the corresponding protein of an uncultured virus (MW522971) associated with Altiarchaeota[57], whereas the former protein is most similar to the DNA polymerase encoded by the spindle-shaped virus infecting marine Nitrososphaeria[55] (Table S3). Ga0209977_10002196 and Ga0209976_10004438 are partial genomes that lack the region encompassing the replication modules. Notably, WyrdV1 lacks either the polymerase or the rolling circle endonuclease gene and instead at the equivalent locus contains an array of short genes, some of which might encode replication initiators. Similar dramatic variation in gene content has been previously observed only in haloarchaeal viruses of the family *Pleolipoviridae*, where members of three genera encode non-homologous genome replication proteins[58]. Our present observations further illuminate the remarkable plasticity of the genome replication modules and genome structures in relatively closely related archaeal viruses, in general, and in wyrdviruses, in particular.

Enigmatic MGEs and auxiliary genes of asgardarchaeal viruses

In addition to the three groups of viruses, we identified seven other contigs (7H_11, 8H_18, 8H_67, 10H_0, 7H_42, Ga0114923_10000127 and Ga0209976_10000148) targeted by asgardarchaeal CRISPR spacers (Table S3). Four of these contigs, 7H_11, 8H_18, 10H_0 and Ga0114923_10000127, were assembled as circular molecules and likely represent complete MGE genomes of 8,776 bp, 8,776 bp, 84,544 bp and 58,806 bp, respectively, whereas Ga0209976_10000148 (48,997 bp), 7H_42 (44,162 bp) and 8H_67 (13,282 bp) appeared to be partial. The seven contigs do not encode identifiable homologs of major capsid proteins of known viruses and are likely to represent either unknown viruses or non-viral MGEs, such as plasmids (see Supplementary text for description of the asgardarchaeal MGEs).

Some of the asgardarchaeal MGEs and viruses encode auxiliary functions, including metabolic genes, such as phosphoadenosine phosphosulfate (PAPS) reductase, which could facilitate sulfur metabolism and/or synthesis of sulfur-containing amino acids[59]; enzymes involved in nucleotide metabolism, including dUTPase, thymidylate synthase X and nucleoside pyrophosphohydrolase MazG, which might function in disarming antiviral systems triggered by nucleotide-based alarmones, such as ppGpp (for more details see Supplementary text).

Discussion

Here we describe three previously undetected, distinct groups of viruses associated with asgardarchaea of the lineages Lokiarchaeia and Thorarchaeia. Each group was identified in marine sediment samples from geographically remote sites (Fig. 6), suggesting a wide distribution of these viruses in asgardarchaea inhabited ecosystems. In addition, we recovered seven CRISPR-targeted MGEs associated with Lokiarchaeia, Thorarchaeia and Heimdallarchaeia that might represent distinct viruses with structural and morphogenetic proteins unrelated to those of any known viruses or, perhaps, more likely, plasmids. Although verdandiviruses, skuldviruses and wyrdviruses, in all likelihood, do not comprise the complete Asgard virome, they provide important insights into the diversity and evolution of the Asgard viruses. All three virus groups are sufficiently distinct from previously characterized viruses to be considered as founding representatives of three previously undescribed families. Wyrdviruses appear to be evolutionarily related to spindle-shaped viruses and thus belong to one of the archaea-specific groups of viruses[19], not known in bacteria or eukaryotes. In archaea, spindle-shaped viruses are widely distributed and infect hyperthermophilic, halophilic and ammonia-oxidizing hosts from different phyla[48,55]. Thus, wyrdviruses might further expand the reach of spindle-shaped viruses to asgardarchaea, supporting the notion that this group of viruses was associated with the last archaeal common ancestor (LACA)[60]. By contrast, verdandiviruses and skuldviruses encode HK97-like and DJR MCPs, respectively, and accordingly, at the highest taxonomic level, belong to the realms *Duplodnaviria* and *Varidnaviria*. Viruses of both realms have deep evolutionary origins and were proposed to have been present in both LUCA and LACA[60]. The current work further supports this inference. Although members of both *Duplodnaviria* and *Varidnaviria* also infect eukaryotes, analyses of the verdandivirus and skuldvirus genome and protein sequences unequivocally show that they are more closely, even if distantly, related to their respective

prokaryotic relatives. Thus, no putative direct ancestors of eukaryotic viruses were detected. Further exploration of the asgardarchaeal virome is needed to determine whether any of the virus groups associated with extant eukaryotes originate from Asgard viruses.

All prokaryotic members of the realms *Duplodnaviria* and *Varidnaviria* are lytic viruses, which are released from the host cells by lysis (although some alternate lysis with lysogeny)[61-63]. Thus, verdandiviruses and skuldviruses are also likely to kill their hosts at the end of the infection cycle, thereby promoting the turnover of asgardarchaea and nutrient cycling in deep-sea ecosystems. This possibility is consistent with previous results showing that viruses in deep-sea sediments lyse archaea faster compared to bacteria[64]. By contrast, the mechanism of virion release employed by spindle-shaped viruses does not involve cell lysis, with virions continuously released from chronically infected cells[55,56]. Infection of marine Nitrososphaeria with spindle-shaped virus NSV1resulted in inhibition of the host growth and was accompanied by severe reduction in the rate of ammonia oxidation and nitrite reduction[55]. Thus, infection dynamics and the impact of wyrdviruses is likely to be quite different from those of verdandiviruses and skuldviruses. Finally, some of the asgardarchaeal viruses carry auxiliary metabolic genes, such as the gene encoding PAPS reductase, which might boost the metabolism of infected cells.

The present work is accompanied by two other studies, by Tamarit et al[65] and Rambo et al[66], respectively, describing the identification of other asgardarchaeal viruses. Rambo et al describe several distinct groups of viruses, all members of the class *Caudoviricetes*. The complete genomes of these viruses are considerably larger than those of verdandiviruses described here and have distinct gene contents, justifying their placements into separate families. The study by Tamarit et al describes two viruses, Huginnvirus and Muninnvirus, related to icosahedral and spindle-shaped viruses, respectively. Huginnvirus is only distantly related to skuldviruses described here, with the two groups of viruses forming disconnected clusters in the MCP network (Fig. 4b). Muninnvirus is distantly related to wyrdviruses associated with Lokiarchaeia from our study, but infects a host from the asgardarchaeal class Odinarchaeia. Thus, viruses described in the three studies complement each other by representing distinct virus groups. Collectively, these findings provide valuable insights into the virome of asgardarchaea and open the door for understanding virus-host interactions in the deep-sea ecosystems. Undoubtedly, many more asgardarchaeal virus groups remain to be discovered, which should clarify the contribution of these viruses to the evolution of the extant eukaryotic virome.

Methods

Site description and sampling

A total of 365 m of sediment cores were recovered from Hole C9001 C of Site C9001 (water depth of 1,180 m) located at the forearc basin off the Shimokita Peninsula, Japan (41˚10.638N, 142˚12.081E) by the drilling vessel Chikyu during JAMSTEC CK06-06 cruise in 2006. Coring procedure, subsampling for the molecular analyses, profiles of lithology, age model, porewater inorganic chemistry, organic chemistry and cell abundance, summary of the molecular microbiology in sediments at site C9001 were reported previously [23,67] (and references therein). Subsampled whole round cores (WRCs) for microbiology were stored at -80˚C.

Sample descriptions

A total of 12 sediment samples at depths of 0.9, 9.3, 18.5, 30.8, 48.3, 59.5, 68.8, 87.7, 116.4, 154.3, 254.7 and 363.3 mbsf were used in this study. These representatives were chosen based on porewater chemical profiles, lithostratigraphic properties, molecular biology data on the core sediments and F430 contents as follows. Note that 10 samples at depths of 0.9, 9.3, 18.5, 30.8, 48.3, 59.5, 116.4, 154.3, 254.3 and 363.3 mbsf were previously analyzed by SSU rRNA gene tag sequencing with a 454 FLX Titanium sequencer, and details have been reported in ref [23].

At 0.9 mbsf, the uppermost section of the core column was chosen. At 9.3 mbsf, the region just beneath the SMTZ was chosen. At 18.5 mbsf, highest relative abundance of Lokiarchaeia was found in the previous SSU rRNA gene tag sequencing. At 48.3 mbsf, relatively higher abundance of SAGMEG in the archaeal community was observed. At 59.5 mbsf predominance of a lineage in Nanoarchaeia (formerly Woesearchaeota) was observed. For the sediment, at depth of 68.8 mbsf, the sample harbored anomalously high F430 concentrations, and the sample from a depth of 87.7 mbsf was used as a reference. Sediment at 116.4 mbsf consisted of ash/pumice, whereas other samples used in this study were pelagic clay sediments, and predominance of Bathyarchaeia was observed.

Samples from 154.3, 254.3 and 363.3 mbsf were selected in 100 m depth intervals from the bottom of this drilling hole [23,67].

Shotgun metagenomics and SSU rRNA gene tag sequencing

DNA extraction and shotgun metagenome library construction were described in ref [68]. Briefly, total environmental DNA was extracted from approximately 5 g of sediment subsampled from inner part of WRC using DNeasy PowerMax Soil Kit (Qiagen) with a minor modification [69]. Then, further purification was performed, NucleoSpin gDNA Clean-up (MACHERY-NAGEL, Germany). The purified DNA was used for Illumina sequencing library construction using a KAPA Hyper Prep Kit (for Illumina) (KAPA Biosystems, Wilmington, MA, USA). Sequencing was performed on an Illumina HiSeq 2500 platform (San Diego, CA, USA) and 250-bp paired-end reads were generated.

Assembly, binning, and Asgardarchaeota phylogeny

Contigs from each sample were assembled using MetaSpades v0.7.12-r1039 [70] and binned with UniteM (https://github.com/dparks1134/unitem).
Metagenome-assembled genome (MAG) sets were generated after a dereplication of bins using DAS_Tool v1.1.2 [71] and quality check using CheckM [72]. MAGs with an estimated quality (completeness - 4*contamination) of less than 30% were excluded in the downstream analysis. Taxonomic assignments of all MAGs were performed using the 'classify' function in GTDBtk [73].

The phylogenetic analysis included 20 Asgardarchaeota MAGs, together with 255 published Asgardarchaeota and 64 non-Asgardarchaeota archaeal genomes. A marker set consisting of 53 ribosomal proteins of the 239 genomes were identified and independently aligned using the 'identify' and 'align' functions in GTDBtk [73]. The individual multiple sequence alignments were then concatenated and trimmed in GTDBtk using 'trim_msa' (--min_perc_aa 0.4). Maximum likelihood phylogenies of Asgardarchaeota MAGs were initially estimated by FastTree v2.1.11 [74] with default settings, and subsequently inferred using IQ-Tree v1.6.984 [75] under the LG+C10+F+G+PMSF model with 100 bootstraps. The final consensus tree was visualized and beautified in iTOL [76].

Assembly and analysis of viral contigs

Potential viral and MGE contigs were assembled from metagenomic reads by MetaViralSPAdes pipeline with default parameters [77]. Direct and inverted terminal repeats were identified by BLASTN [78]. Open reading frames were identified with Prokka [79] and annotated using HHsearch [80] against Pfam, PDB, SCOPe, CDD and viral protein sequence databases. Minimum information about uncultured virus genomes (MIUViG) [81] assembled in the course of this work is provided in Table S7. Read depth along the assembled viral genomes is shown in Fig. S4. To identify similar contigs in public databases we searched sequences of major capsid proteins in GenBank and JGI databases (BLASTP, E-value cutoff of 1e-05, 70% query coverage). Genomes of assembled viral genomes were compared and visualized using EasyFig [82] with the tBLASTx option. Network analysis of viral genomes was performed using vConTACT v2.0.9.19 with default parameters against Viral RefSeq v201 reference database [83]. The virus network was visualized with Cytoscape [84].

Collection of the CRISPR spacer dataset

CRISPR arrays were detected in metagenomic contigs and 255 published Asgardarchaeota MAGs using minced v0.4.2 [85] and CRISPRDetect [86]. CRISPR arrays from metagenomic contigs were assigned to “Asgard” and “non-Asgard” groups based on CRISPR repeat similarity to previously characterized Asgard CRISPR repeats with 90% identity cutoff over full length of repeat[21]. CRISPR repeat sequences from metagenomic contigs and from Asgard MAGs were then clustered together using all-against-all BLASTN search (E-value cutoff of 1e-05, 90% identity, word size 7) and the result of clustering was visualized in Cytoscape [84]. Consensus sequences for the 9 major clusters of CRISPR repeats were constructed using python package Logomaker [87].

Assignment of CRISPR arrays to Asgardarchaeota

To assemble the CRISPR spacer database, we relied on two categories of CRISPR loci: 1) “reference” sequences of manually curated Asgard CRISPR loci from Makarova et al [21] and CRISPR loci from asgardarchaeal MAGs; 2) “metagenomic” CRISPR loci from contigs originating from the Shimokita Peninsula subseafloor sediments. CRISPR loci from the latter category were assigned to asgardarchaea if the

identity of CRISPR repeat to the reference asgardarchaeal CRISPRs from the former category was above 90% and protein sequences encoded on these CRISPR-containing contigs (when present) had best BLASTP hit to asgardarchaeal proteins. All "reference" contigs were downloaded from GenBank without annotation. We used prodigal and CRISPR-Cas++ to identify *cas* genes in the contigs with CRISPR arrays. Protein sequences were compared to the nr database with BLASTP to confirm Asgard assignment of the contig with CRISPR.

Verdandiviruses were targeted by spacers from 5 "reference" and 4 "metagenomic" CRISPR loci (Extended Data 1). The length of "reference" contigs varied from 3 to 126 kbp. Five reference contigs and one metagenomic contig contained cas genes, with the most conserved Cas protein (Cas1) being assigned to asgardarchaea (Lokiarachaeia, Helarchaeia) by the best BLASTP hit with 72-74% identity. Skuldviruses were targeted by spacers from 3 "reference" and 3 "metagenomic" CRISPR loci, whereas wyrdviruses were targeted by spacers from 2 "reference" and 3 "metagenomic" CRISPR loci (Extended Data 1). CRISPR repeats of "metagenomic" contigs targeting the three groups of viruses were 97% identical to the CRISPR repeats from the "reference" category [21].

Asgardarchaeal CRISPR repeats

Predicted asgardarchaeal CRISPR arrays contained distinct repeats which could be clustered into 9 groups with 90% identity of sequences within groups (Fig. 2A). Comparison of the asgardarchaeal CRISPR repeats to those in public RepeatTyper database produced significant results only for group 9 repeat, which was identified as belonging to CRISPR-Cas type I-E. The closest repeat to group 9 was from *Thermobaculum terrenum*, with 2 mismatches (27/29). Repeats from groups 4 partially matched to *Desulfosarcina* (29/37). Other repeat groups showed no matches in CRISPR-Cas++ database (https://crisprcas.i2bc.paris-saclay.fr/).

Structural modeling and network analysis of the major capsid proteins

Structural modeling of the representative verdnadivirus and skuldvirus major capsid proteins was performed with RoseTTAFold [28]. The MCPs of skuldviruses were compared to the DJR MCPs of other known prokaryotic viruses using CLANS [41], an implementation of the Fruchterman-Reingold force-directed layout algorithm, which treats protein sequences as point masses in a virtual multidimensional space, in which they attract or repel each other based on the strength of their pairwise similarities (CLANS p-values). The reference dataset of DJR MCP was obtained from Ref [42]. Sequences were clustered using CLANS with BLASTP option (E-value of 1e-04) [41].

Phylogenetic analysis of viral proteins

Sequences were aligned using MAFFT in 'Auto' mode [88]. For phylogenetic analysis, uninformative positions we removed using TrimAl with gap threshold of 0.2 for the verdandiviral MCPs and gappyout option for Topo mini-A [89]. The final alignments contained 323 and 277 positions, respectively. The maximum likelihood trees was inferred using IQ-TREE v2 [75]. The best-fitting substitution models were selected by IQ-TREE and were LG+G4 and LG+R5 for verdandiviral MCPs and Topo mini-A, respectively. The trees were visualized with iTOL [76]. Phylogenetic trees and underlying alignments in editable format are provided in Supplementary data 2.

CRISPR-based virus-host assignments

Spacer sequences from asgardarchaeal CRISPR arrays were matched to metagenomic contigs and published Asgard MAGs (BLASTN, E-value 1e-05, 90% identity, word size 7), and viral genomes available at IGV database (default parameters) at https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=WorkspaceBlast&page=viralform. To assess the false positive rate of BLASTN, we used a control set of viral contigs from human metagenomes (~400,000 sequences) [90-93]. Seven spacers out of 2532 (FPR = 0.00276) were matched to a control dataset by BLASTN with the same parameters (E-value 1e-05, 90% identity, word size 7). For each spacer we calculated the difference between identity of best match in the test and control datasets. Spacers with less than 10% identity difference (n=5) were removed as potential low-complexity spacers. SpacePHARER [94] with FDR rate of < 0.001 was used for sensitive search of protospacers with lower nucleotide identity to support the BLASTN results. CRISPRHost (https://github.com/davidchyou/CRISPRHost) and CRISPR Spacer Database and Exploration Tool [95] were used to match large collections of prokaryotic spacers to identified Asgard MGE sequences.

Kmer-based virus-host assignments

WIsH [96] and PHIST [97], kmer-based methods which compare 8-mer (WIsH) and 25-mer (PHIST) nucleotide frequencies of viruses and potential hosts, were applied to validate prediction of CRISPR-based virus-host assignment. Unlike other kmer-based methods, PHIST and WIsH can predict virus-host interactions for user-provided host sequences, rather than pre-computed host databases. For the potential host dataset, we combined 195 MAGs of Asgardarchaeota, 267 non-Asgard MAGs reconstructed from Shimokita Peninsula metagenomes, 2143 RefSeq archaeal genomes and 1077 representative bacterial genomes (Table S8). For the virus dataset, we used MGE sequences targeted by Asgard spacers. Viral contigs from human metagenomes were used as a negative dataset to calculate null parameters for models of hosts for WIsH.

WIsH confirmed asgardarchaeal hosts for all MGE sequences targeted by asgardarchaeal spacers with a P-value ranging from 1.18e-06 for WyrdV1 to 1.32e-01 for VerdaV2 (Table S5). In the 25-mer-based PHIST analysis, we only considered predictions based on >10 kmers. Using these parameters, hosts were predicted for 17 out of 28 (60%) contigs. Of these, 11 (65%) contigs, representing all groups of viruses and MGEs described in this work, were predicted to be associated with asgardarchaea. However, the remaining 6, all verdandiviruses, were affiliated to Thermoplasmatota MAG 2H_mb2_bin16 assembled from Shimokita peninsula dataset. We note that the latter prediction likely results from the inclusion of a small (~2 kbp) verdandivirus-like contig into Thermoplasmatota MAG 2H_mb2_bin16, which is likely a binning artefact, because full-length verdandivirus contig was assigned to Lokiarachaeia MAG. Indeed, we could not confirm Thermoplasmatota assignment as a verdandivirus host with CRISPR spacer analysis.

To further verify the validity of our CRISPR matches, we checked the predicted asgardarchaeal viruses and MGEs for the presence of non-Asgard protospacers using large spacer collections of CRISPRHost and CRISPR Spacer Database and Exploration Tool (CRISPR_SDET). CRISPRHost did not find any protospacers with the default parameters. CRISPR_SDET search was performed with 90% identity threshold. Two protospacers were found: verdandivirus VerdaV4 was assigned to *Floccifex porci*, a Firmicutes bacterium from pig gut metagenome; two mobile elements from the "Other" group, JGI-127 and JGI-148, were assigned to *Treponema* species, a pathogenic bacterium of humans and other mammals. Given that neither of these bacteria reside in deep-sea sediments, the two spacers are likely to represent false positives.

Code Availability

No custom code was used.

Data availability

The raw reads as well as assembled virus and MGE genome sequences from the metagenomes described in this study are available at NCBI under BioProject PRJDB12054, BioSample accessions: SAMD00394285-SAMD00394296. Accession numbers of the 20 MAGs assembled in the course of this study are listed in Supplementary table 1.

Intramural Research Program of the National Institutes of Health of the USA (National Library of Medicine). The work by C.R and J. S. is funded by the Australian Research Council (ARC) Future Fellow Award (FT170100213) awarded to CR. T.N. was partly supported by MEXT KAKENHI Grant Number JP19H05684 within JP19H05679 (Post-Koch Ecology) and 16H06429, 16K21723 and 16H06437 (NeoVirology).

Author Contributions Statement

M.K. initiated the project and designed research. T.N. collected the samples, extracted and sequenced the DNA. J.S. and C.R. assembled, curated and analyzed the asgardarchaeal MAGs. S.M. assembled the

asgardarchaeal CRISPR and viral datasets. S.M., N.Y., E.V.K. and M.K. analyzed the viral sequences. S.M. and M.K. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

References

1 Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553-557, doi:10.1038/s41586-021-03494-3 (2021).

2 Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nat Commun* **9**, 4999, doi:10.1038/s41467-018-07418-0 (2018).

3 Wong, H. L. *et al.* Disentangling the drivers of functional complexity at the metagenomic level in Shark Bay microbial mat microbiomes. *ISME J* **12**, 2619-2639, doi:10.1038/s41396-018-0208-8 (2018).

4 Liu, Y. *et al.* Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota. *ISME J* **12**, 1021-1031, doi:10.1038/s41396-018-0060-x (2018).

5 Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353-358, doi:10.1038/nature21031 (2017).

6 Seitz, K. W., Lazar, C. S., Hinrichs, K. U., Teske, A. P. & Baker, B. J. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J* **10**, 1696-1705, doi:10.1038/ismej.2015.233 (2016).

7 Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173-179, doi:10.1038/nature14447 (2015).

8 Sun, J. *et al.* Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota lineages. *ISME Commun* **1**, 30 (2021).

9 Seitz, K. W. *et al.* Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat Commun* **10**, 1822, doi:10.1038/s41467-019-09364-x (2019).

10 Farag, I. F., Zhao, R. & Biddle, J. F. "Sifarchaeota," a novel Asgard phylum from Costa Rican sediment capable of polysaccharide degradation and anaerobic methylotrophy. *Appl Environ Microbiol* **87**, e02584-02520, doi:10.1128/AEM.02584-20 (2021).

11 Zhang, J. W. *et al.* Newly discovered Asgard archaea Hermodarchaeota potentially degrade alkanes and aromatics via alkyl/benzyl-succinate synthase and benzoyl-CoA pathway. *ISME J* **15**, 1826-1843, doi:10.1038/s41396-020-00890-x (2021).

12 Cai, M. *et al.* Diverse Asgard archaea including the novel phylum Gerdarchaeota participate in organic matter degradation. *Sci China Life Sci* **63**, 886-897, doi:10.1007/s11427-020-1679-1 (2020).

13 Rinke, C. *et al.* A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6**, 946-959, doi:10.1038/s41564-021-00918-8 (2021).

14 Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519-525, doi:10.1038/s41586-019-1916-6 (2020).

15 Lopez-Garcia, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655-667, doi:10.1038/s41564-020-0710-4 (2020).

16 Da Cunha, V., Gaia, M., Nasir, A. & Forterre, P. Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet* **14**, e1007215, doi:10.1371/journal.pgen.1007215 (2018).

17 Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A. & Forterre, P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet* **13**, e1006810, doi:10.1371/journal.pgen.1006810 (2017).

18 Baquero, D. P. *et al.* Structure and assembly of archaeal viruses. *Adv Virus Res* **108**, 127-164, doi:10.1016/bs.aivir.2020.09.004 (2020).

19 Prangishvili, D. *et al.* The enigmatic archaeal virosphere. *Nat Rev Microbiol* **15**, 724-739, doi:10.1038/nrmicro.2017.125 (2017).

20 Dellas, N., Snyder, J. C., Bolduc, B. & Young, M. J. Archaeal Viruses: Diversity, Replication, and Structure. *Annu Rev Virol* **1**, 399-426, doi:10.1146/annurev-virology-031413-085357 (2014).

21 Makarova, K. S. *et al.* Unprecedented diversity of unique CRISPR-Cas-related systems and Cas1 homologs in Asgard archaea. *CRISPR J* **3**, 156-163, doi:10.1089/crispr.2020.0012 (2020).

22 Coclet, C. & Roux, S. Global overview and major challenges of host prediction methods for uncultivated phages. *Curr Opin Virol* **49**, 117-126, doi:10.1016/j.coviro.2021.05.003 (2021).

23 Nunoura, T. *et al.* Variance and potential niche separation of microbial communities in subseafloor sediments off Shimokita Peninsula, Japan. *Environ Microbiol* **18**, 1889-1906, doi:10.1111/1462-2920.13096 (2016).

24 Glass, J. B. *et al.* Microbial metabolism and adaptations in Atribacteria-dominated methane hydrate sediments. *Environ Microbiol* **23**, 4646-4660, doi:10.1111/1462-2920.15656 (2021).

25 Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nat Rev Microbiol* **18**, 125-138, doi:10.1038/s41579-019-0311-5 (2020).

26 Iranzo, J., Krupovic, M. & Koonin, E. V. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *mBio* **7**, e00978-00916, doi:10.1128/mBio.00978-16 (2016).

27 Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**, doi:10.1128/MMBR.00061-19 (2020).

28 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876, doi:10.1126/science.abj8754 (2021).

29 Wang, Z. *et al.* Structure of the Marine Siphovirus TW1: Evolution of Capsid-Stabilizing Proteins and Tail Spikes. *Structure* **26**, 238-248 e233, doi:10.1016/j.str.2017.12.001 (2018).

30 Hendrix, R. W. Tail length determination in double-stranded DNA bacteriophages. *Curr Top Microbiol Immunol* **136**, 21-29, doi:10.1007/978-3-642-73115-0_2 (1988).

31 Mahony, J. *et al.* Functional and structural dissection of the tape measure protein of lactococcal phage TP901-1. *Sci Rep* **6**, 36667, doi:10.1038/srep36667 (2016).

32 Pope, W. H. *et al.* Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *Elife* **4**, e06416, doi:10.7554/eLife.06416 (2015).

33 Krupovic, M., Forterre, P. & Bamford, D. H. Comparative analysis of the mosaic genomes of tailed archaeal viruses and

proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. *J Mol Biol* **397**, 144-160, doi:10.1016/j.jmb.2010.01.037 (2010).
34 Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D. & Koonin, E. V. Viruses of archaea: Structural, functional, environmental and evolutionary genomics. *Virus Res* **244**, 181-193, doi:10.1016/j.virusres.2017.11.025 (2018).
35 Liu, Y. *et al.* Diversity, taxonomy, and evolution of archaeal viruses of the class *Caudoviricetes*. *PLoS Biol* **19**, e3001442, doi:10.1371/journal.pbio.3001442 (2021).
36 Gardner, A. F., Bell, S. D., White, M. F., Prangishvili, D. & Krupovic, M. Protein-protein interactions leading to recruitment of the host DNA sliding clamp by the hyperthermophilic Sulfolobus islandicus rod-shaped virus 2. *J Virol* **88**, 7105-7108, doi:10.1128/JVI.00636-14 (2014).
37 Gussow, A. B. *et al.* Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nat Commun* **11**, 3784, doi:10.1038/s41467-020-17652-0 (2020).
38 Li, Y. & Bondy-Denomy, J. Anti-CRISPRs go viral: The infection biology of CRISPR-Cas inhibitors. *Cell Host Microbe* **29**, 704-714, doi:10.1016/j.chom.2020.12.007 (2021).
39 Krupovic, M. & Bamford, D. H. Virus evolution: how far does the double beta-barrel viral lineage extend? *Nat Rev Microbiol* **6**, 941-948, doi:10.1038/nrmicro2033 (2008).
40 Hong, C. *et al.* A structural model of the genome packaging process in a membrane-containing double stranded DNA virus. *PLoS Biol* **12**, e1002024, doi:10.1371/journal.pbio.1002024 (2014).
41 Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702-3704, doi:10.1093/bioinformatics/bth444 (2004).
42 Yutin, N., Bäckström, D., Ettema, T. J. G., Krupovic, M. & Koonin, E. V. Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. *Virol J* **15**, 67, doi:10.1186/s12985-018-0974-y (2018).
43 Abrescia, N. G. *et al.* Insights into virus evolution and membrane biogenesis from the structure of the marine lipid-containing bacteriophage PM2. *Mol Cell* **31**, 749-761, doi:10.1016/j.molcel.2008.06.026 (2008).
44 Oksanen, H. M. & ICTV Report Consortium. ICTV Virus Taxonomy Profile: *Corticoviridae*. *J Gen Virol* **98**, 888-889, doi:10.1099/jgv.0.000795 (2017).
45 Krupovic, M. & Bamford, D. H. Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Genomics* **8**, 236, doi:10.1186/1471-2164-8-236 (2007).
46 Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat Commun* **10**, 3425, doi:10.1038/s41467-019-11433-0 (2019).
47 Takahashi, T. S. *et al.* Expanding the type IIB DNA topoisomerase family: identification of new topoisomerase and topoisomerase-like proteins in mobile genetic elements. *NAR Genom Bioinform* **2**, lqz021, doi:10.1093/nargab/lqz021 (2020).
48 Krupovic, M., Quemin, E. R., Bamford, D. H., Forterre, P. & Prangishvili, D. Unification of the globally distributed spindle-shaped viruses of the Archaea. *J Virol* **88**, 2354-2358, doi:10.1128/JVI.02941-13 (2014).
49 Bath, C., Cukalac, T., Porter, K. & Dyall-Smith, M. L. His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. *Virology* **350**, 228-239, doi:10.1016/j.virol.2006.02.005 (2006).
50 Wang, F. *et al.* Spindle-shaped archaeal viruses evolved from rod-shaped ancestors to package a larger genome. *Cell* **185**, 1297-1307.e1211, doi:10.1016/j.cell.2022.02.019 (2022).
51 Hong, C. *et al.* Lemon-shaped halo archaeal virus His1 with uniform tail but variable capsid structure. *Proc Natl Acad Sci U S A* **112**, 2449-2454, doi:10.1073/pnas.1425008112 (2015).
52 Quemin, E. R. *et al.* Sulfolobus Spindle-Shaped Virus 1 Contains Glycosylated Capsid Proteins, a Cellular Chromatin Protein, and Host-Derived Lipids. *J Virol* **89**, 11681-11691, doi:10.1128/JVI.02270-15 (2015).
53 Roux, S. *et al.* Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol* **4**, 1895-1906, doi:10.1038/s41564-019-0510-x (2019).
54 Straus, S. K. & Bo, H. E. Filamentous bacteriophage proteins and assembly. *Subcell Biochem* **88**, 261-279, doi:10.1007/978-981-10-8456-0_12 (2018).
55 Kim, J. G. *et al.* Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. *Proc Natl Acad Sci U S A* **116**, 15645-15650, doi:10.1073/pnas.1905682116 (2019).
56 Quemin, E. R. *et al.* Eukaryotic-like virus budding in Archaea. *mBio* **7**, e01439-01416, doi:10.1128/mBio.01439-16 (2016).
57 Rahlff, J. *et al.* Lytic archaeal viruses infect abundant primary producers in Earth's crust. *Nat Commun* **12**, 4642, doi:10.1038/s41467-021-24803-4 (2020).
58 Bamford, D. H. *et al.* ICTV Virus Taxonomy Profile: *Pleolipoviridae*. *J Gen Virol* **98**, 2916-2917, doi:10.1099/jgv.0.000972 (2017).
59 Summer, E. J., Gill, J. J., Upton, C., Gonzalez, C. F. & Young, R. Role of phages in the pathogenesis of Burkholderia, or 'Where are the toxin genes in Burkholderia phages?'. *Curr Opin Microbiol* **10**, 410-417, doi:10.1016/j.mib.2007.05.016 (2007).
60 Krupovic, M., Dolja, V. V. & Koonin, E. V. The LUCA and its complex virome. *Nat Rev Microbiol* **18**, 661-670, doi:10.1038/s41579-020-0408-x (2020).
61 Cahill, J. & Young, R. Phage Lysis: Multiple Genes for Multiple Barriers. *Adv Virus Res* **103**, 33-70, doi:10.1016/bs.aivir.2018.09.003 (2019).
62 Snyder, J. C. & Young, M. J. Lytic viruses infecting organisms from the three domains of life. *Biochem Soc Trans* **41**, 309-313, doi:10.1042/BST20120326 (2013).
63 Krupovic, M., Daugelavicius, R. & Bamford, D. H. A novel lysis system in PM2, a lipid-containing marine double-stranded DNA bacteriophage. *Mol Microbiol* **64**, 1635-1648, doi:10.1111/j.1365-2958.2007.05769.x (2007).
64 Danovaro, R. *et al.* Virus-mediated archaeal hecatomb in the deep seafloor. *Sci Adv* **2**, e1600492, doi:10.1126/sciadv.1600492 (2016).
65 Tamarit, D. *et al.* A closed *Candidatus* Odinarchaeum chromosome exposes Asgard archaeal viruses. *Nat Microbiol* **In press** (2022).

66 Rambo, I. M., de Anda, V., Langwig, M. V. & Baker, B. J. Unique viruses that infect Archaea related to eukaryotes. *BioRxiv* https://doi.org/10.1101/2021.07.29.454249 (2021).

67 Kaneko, M. *et al.* Insights into the methanogenic population and potential in subsurface marine sediments based on coenzyme F430 as a function-specific compound analysis. *JACS Au* **1**, 1743-1751, doi:10.1021/jacsau.1c00307 (2021).

68 Hirai, M. *et al.* Library construction from subnanogram DNA for pelagic sea water and deep-sea sediments. *Microbes Environ* **32**, 336-343, doi:10.1264/jsme2.ME17132 (2017).

69 Hiraoka, S. *et al.* Microbial community and geochemical analyses of trans-trench sediments for understanding the roles of hadal environments. *ISME J* **14**, 740-756, doi:10.1038/s41396-019-0564-z (2020).

70 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824-834, doi:10.1101/gr.213959.116 (2017).

71 Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* **3**, 836-843, doi:10.1038/s41564-018-0171-1 (2018).

72 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

73 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, doi:10.1093/bioinformatics/btz848 (2019).

74 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

75 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).

76 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293-W296, doi:10.1093/nar/gkab301 (2021).

77 Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126-4129, doi:10.1093/bioinformatics/btaa490 (2020).

78 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

79 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

80 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473, doi:10.1186/s12859-019-3019-7 (2019).

81 Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat Biotechnol* **37**, 29-37, doi:10.1038/nbt.4306 (2019).

82 Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**, 1009-1010, doi:10.1093/bioinformatics/btr039 (2011).

83 Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**, 632-639, doi:10.1038/s41587-019-0100-8 (2019).

84 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

85 Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209, doi:10.1186/1471-2105-8-209 (2007).

86 Biswas, A., Staals, R. H., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356, doi:10.1186/s12864-016-2627-0 (2016).

87 Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272-2274, doi:10.1093/bioinformatics/btz921 (2020).

88 Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* **20**, 1160-1166, doi:10.1093/bib/bbx108 (2019).

89 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).

90 Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* **6**, 960-970, doi:10.1038/s41564-021-00928-6 (2021).

91 Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109 e1099, doi:10.1016/j.cell.2021.01.029 (2021).

92 Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, 724-740 e728, doi:10.1016/j.chom.2020.08.003 (2020).

93 Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-541 e525, doi:10.1016/j.chom.2019.09.009 (2019).

94 Zhang, R. *et al.* SpacePHARER: Sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics*, doi:10.1093/bioinformatics/btab222 (2021).

95 Dion, M. B. *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res* **49**, 3127-3138, doi:10.1093/nar/gkab133 (2021).

96 Galiez, C., Siebert, M., Enault, F., Vincent, J. & Soding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113-3114, doi:10.1093/bioinformatics/btx383 (2017).

97 Zielezinski, A., Deorowicz, S. & Gudys, A. PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics*, doi:10.1093/bioinformatics/btab837 (2021).

Figures

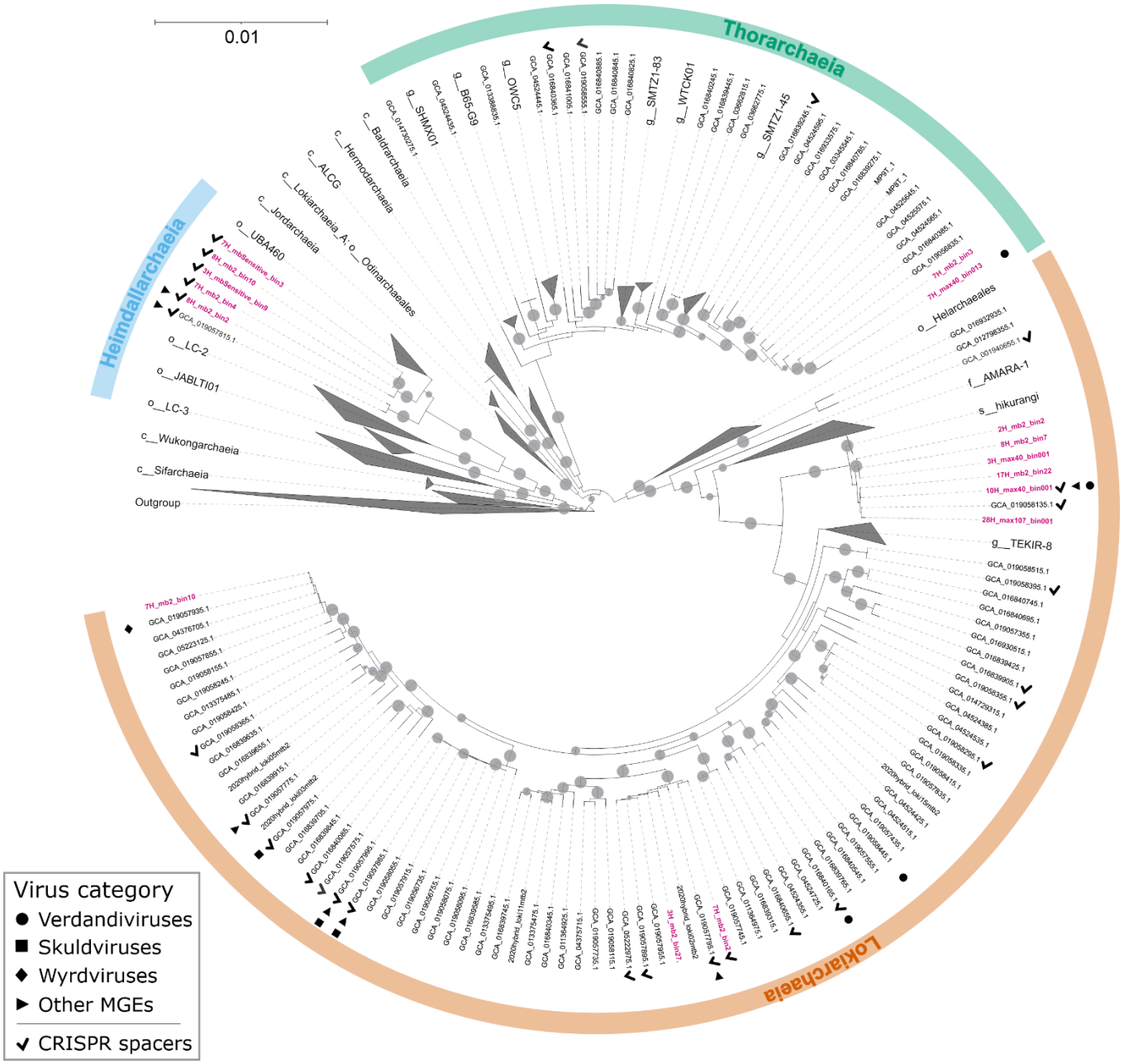


Figure 1 | Phylogenomic tree of Asgardarchaeota. The alignment is based on a concatenated set of 53 protein markers (subsampled to 42 sites each, resulting in a total alignment length of 2058 sites) from 339 taxa. These taxa encompass 276 Asgardarchaeota MAGs, including 16 recovered in this study (highlighted in bold pink) (MAGs 3H_mb2_bin20, 4H_max40_bin02,1 4H_mb2_bin40, and 7H_mb2_bin7 were removed after the alignment trimming step due to low completeness), and 63 non-Asgardarchaeota species representatives from GTDB release RS95 as outgroup. Maximum-likelihood analysis was performed using IQ-TREE under the LG+C10+F+G+PMSF model. The tree is rooted on the non-Asgardarchaeota group. Nodes with bootstrap statistical support over 90% are indicated with grey dots. MAGs containing viral contigs and CRISPR arrays are indicated with different symbols and the key is provided in the bottom left corner. GCA_019058445.1 contains a defective verdandivirus-derived provirus (Extended Data 3).

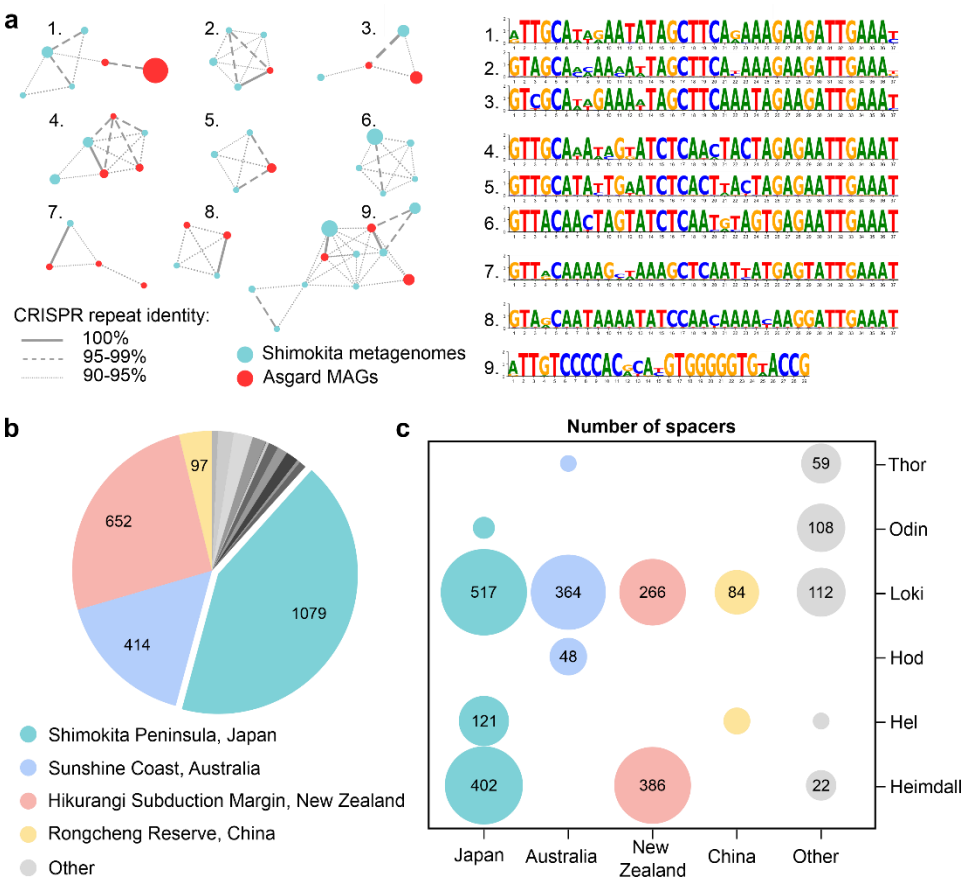


Figure 2 | Description of the Asgardarchaeota CRISPR spacer dataset. a, Similarity of CRISPR repeats from metagenomic data (Bermuda) and CRISPR repeats from Asgardarchaeota MAGs (red). Unique CRISPR repeat sequences are shown as nodes in the network. The diameter of the node is proportional to the number of spacers associated with the repeat sequence in the dataset. Identical CRISPR repeats from metagenomic data and from Asgard MAGs are connected with solid lines. Dashed and dotted lines connect CRISPR repeat sequences with 95-99% and 90-95% identities, respectively. Consensus sequences for the nine major clusters are shown on the right. **b,** The source of the Asgardarchaeota CRISPR spacers. The number of spacers in the dataset originating from different geographical sites is indicated on the pie chart. **c,** Taxonomic distribution of Asgardarchaeota MAGs with CRISPR arrays for different isolation sites. The size of the circle is proportional to the number of spacers in the dataset.

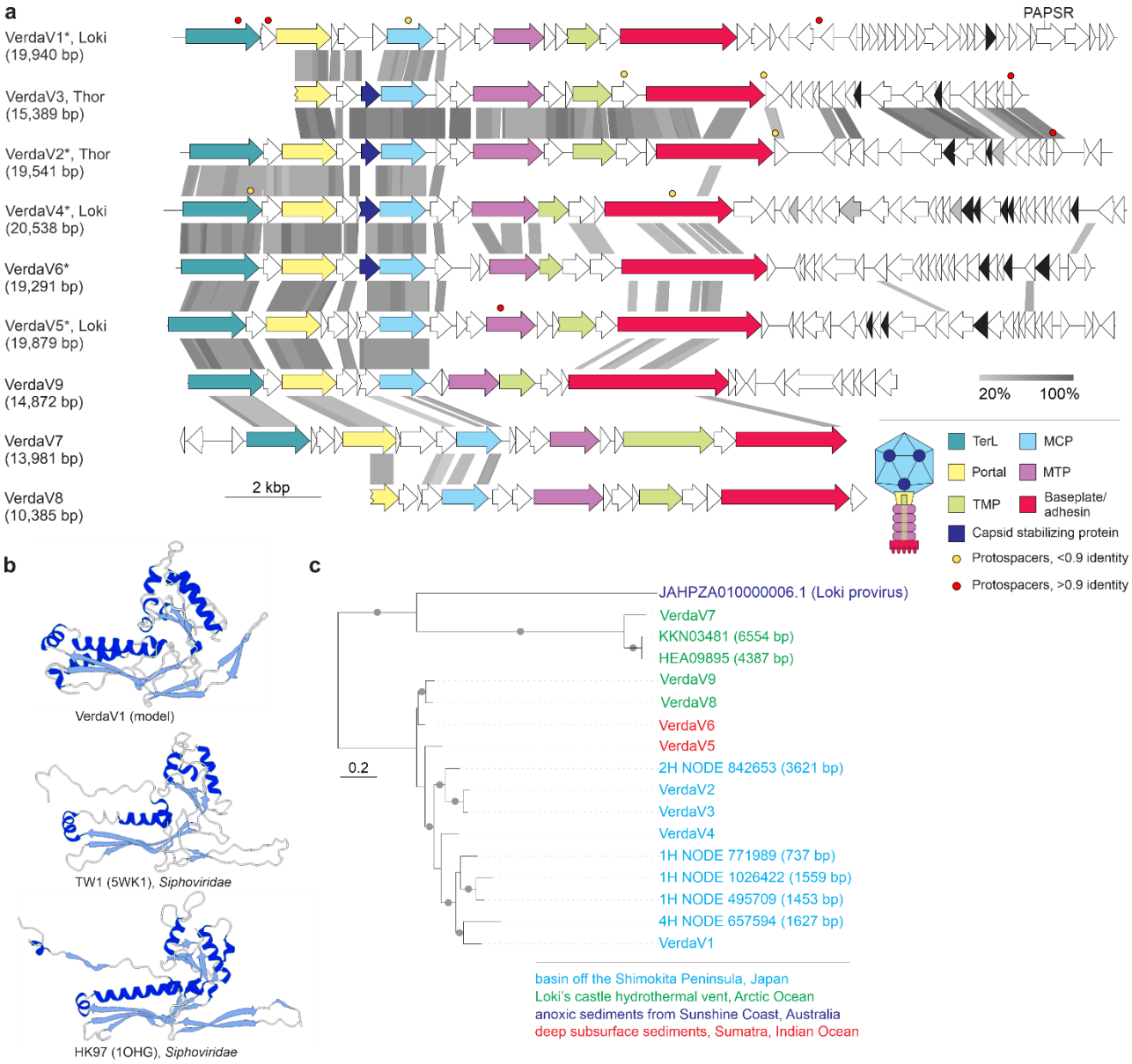


Figure 3 | Diversity of verdandiviruses. a, Genome maps of verdandiviruses. Homologous genes are shown using the same colors and the key is provided at the bottom of the panel. Also shown is the deduced schematic organization of the verdandivirus virion with colors marching those of the genes encoding the corresponding proteins. Genes encoding putative DNA-binding proteins with Zn-binding and helix-turn-helix domains are colored in black and grey, respectively. Colored circles indicate the positions of protospacers. Grey shading connects genes displaying sequence similarity at the protein level, with the percent of sequence identity depicted with different shades of grey (see scale at the bottom). Asterisks denote complete genomes assembled as circular contigs. Abbreviations: PAPSR, phosphoadenosine phosphosulfate reductase; TMP, tail tape measure protein; MTP, major tail protein; MCP, major capsid protein; TerL, large subunit of the terminase. **b,** Comparison of the structural model of the major capsid protein of verdnadivirus AsTV-10H2 with the corresponding structures of siphoviruses TW1 and HK97. The models are colored according to the secondary structure: α-helices, dark blue; β-strands, light blue. **c,** Maximum likelihood phylogeny of the major capsid proteins encoded by verdandiviruses. Taxa are colored based on the source of origin (key at the bottom of the figure), with those for which genomic contigs are shown in panel A shown in bold. The tree was constructed using the automatic optimal model selection (LG+G4) and is mid-point rooted. The scale bar represents the number of substitution per site. Circles at the nodes denote aLRT SH-like branch support values larger than 90%.

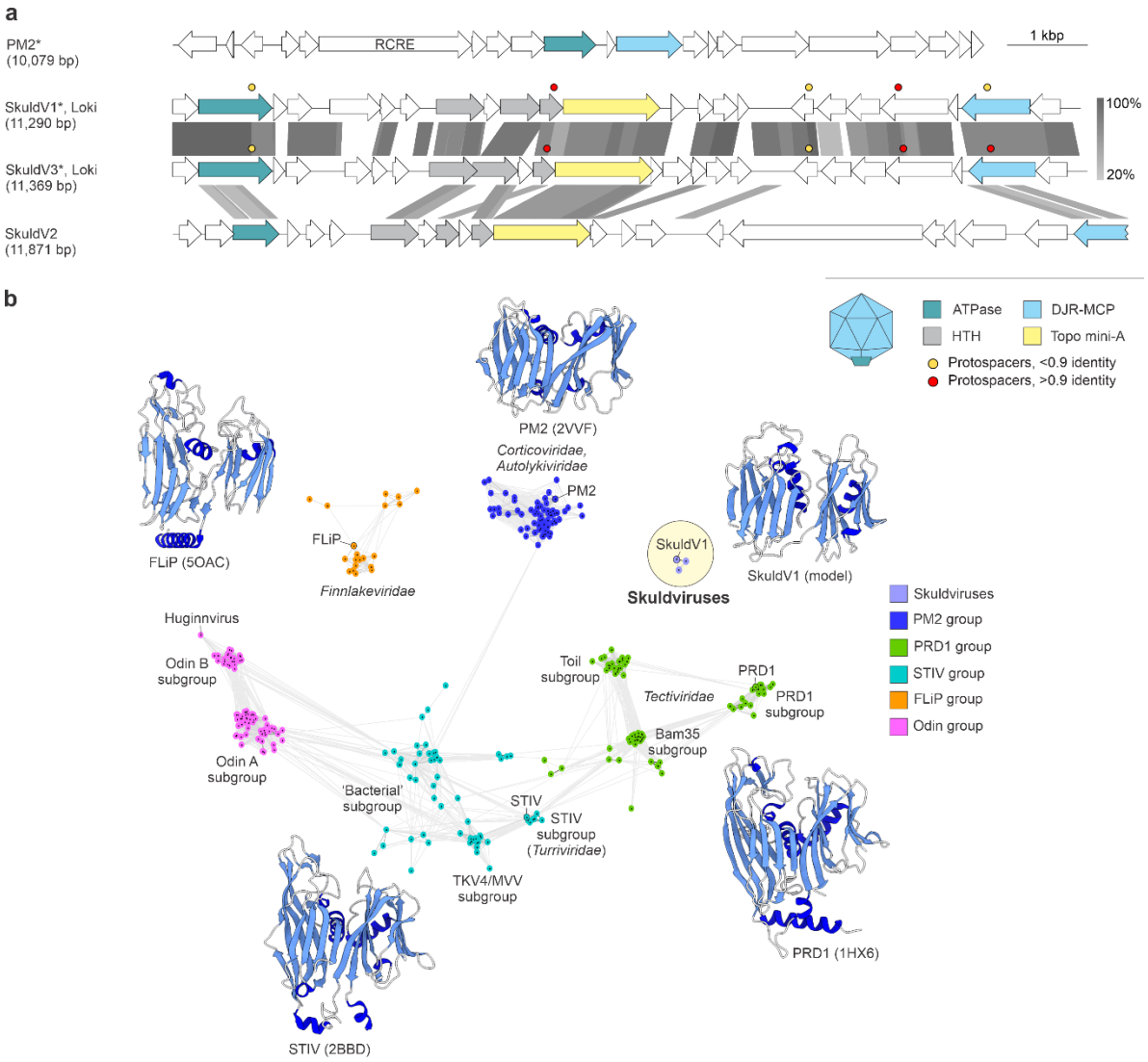


Figure 4 | Diversity of skuldviruses. a, Genome maps of skuldviruses and PM2 virus. Homologous genes are shown using the same colors and the key is provided at the bottom of the panel. Also shown is the deduced schematic organization of the skuldvirus virion with colors marching those of the genes encoding the corresponding proteins. Colored circles indicate the positions of protospacers. Grey shading connects genes displaying sequence similarity at the protein level, with the percent of sequence identity depicted with different shades of grey (see scale on the right). Asterisks denote complete genomes assembled as circular contigs. Abbreviations: RCRE, rolling circle replication endonuclease; DJR-MCP, double jelly-roll major capsid protein; HTH, helix-turn-helix. Genome map of corticovirus PM2 is shown for comparison. **b,** Sequence similarity network of prokaryotic virus DJR MCPs. Protein sequences were clustered by the pairwise sequence similarity using CLANS. Lines connect sequences with CLANS P-value $\leq$ 1e−04. CLANS uses p-values of BLASTP comparisons calculated from Poisson distribution of high scoring segment pairs. Different previously defined groups [42] of DJR MCP are shown as clouds of differentially colored circles, with the key provided on the right. Note that the Odin group has been named as such previously because some of the MCPs in this cluster originated from the Odinarchaeia bin [42]. The position of Huginnvirus described by Tamarit et al [65] is indicated. PRD1, Toil, and Bam35 subgroups are named after the corresponding members of the family *Tectiviridae*. Skuldviruses are highlighted within a yellow circle. When available, MCP structures of viruses representing each group are shown next to the corresponding cluster (PDB accession numbers are given in the parenthesis). Skuldvirus cluster is represented by a structural model of the SkuldV1 MCP. The models are colored according to the secondary structure: α-helices, dark blue; β-strands, light blue.

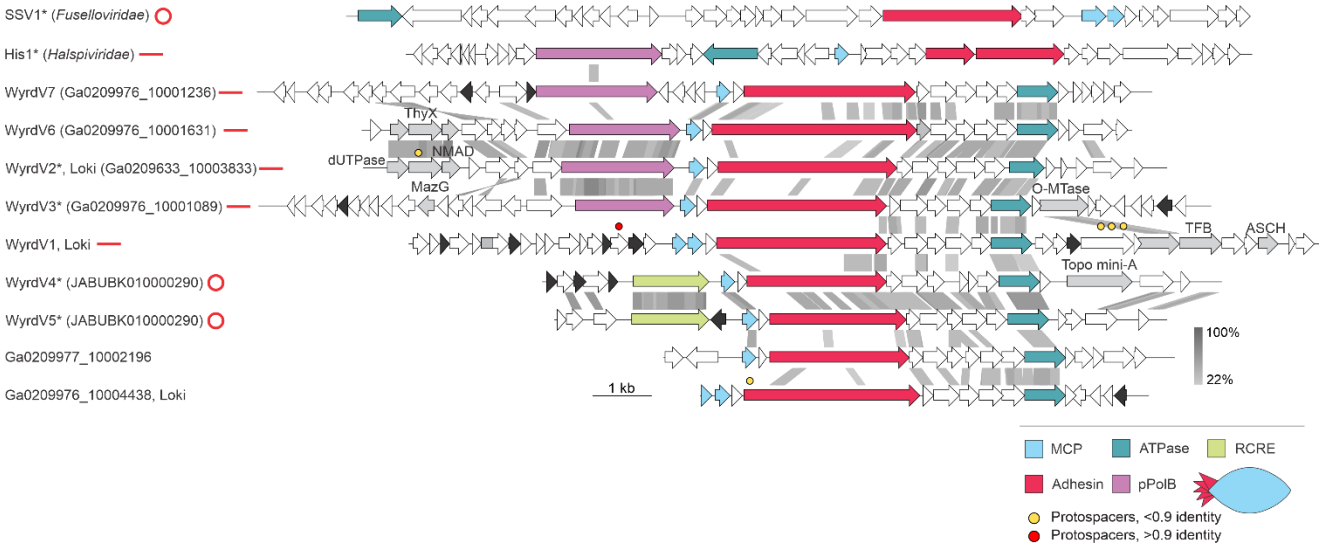


Figure 5 | Diversity of wyrdviruses. Genome maps of wyrdviruses, fusellovirus SSV1 and halspivirus His1. Homologous genes are shown using the same colors and the key is provided at the bottom of the panel. Also shown is the deduced schematic organization of the skuldvirus virion with colors marching those of the genes encoding the corresponding proteins. Colored circles indicate the positions of protospacers. Genes encoding putative DNA-binding proteins with Zn-binding and helix-turn-helix domains are colored in black and grey, respectively. Grey shading connects genes displaying sequence similarity at the protein level, with the percent of sequence identity depicted with different shades of grey (see scale on the right). Asterisks denote complete genomes assembled as circular contigs or linear contigs with terminal inverted repeats. The topology of each genome is shown next to the corresponding name: circle for circular genomes, lines for linear genomes. Abbreviations: ThyX, thymidylate synthase X; NMAD, nucleotide modification associated domain 1 protein; O-MTase, carbohydrate-specific methyltransferase; TFB, transcription initiation factor B; ASCH, ASC-1 homology domain; RCRE, rolling circle replication endonuclease; MCP, major capsid protein, pPolB, protein-primed family B DNA polymerase. Genome maps of SSV1 and His1 viruses are shown for comparison.

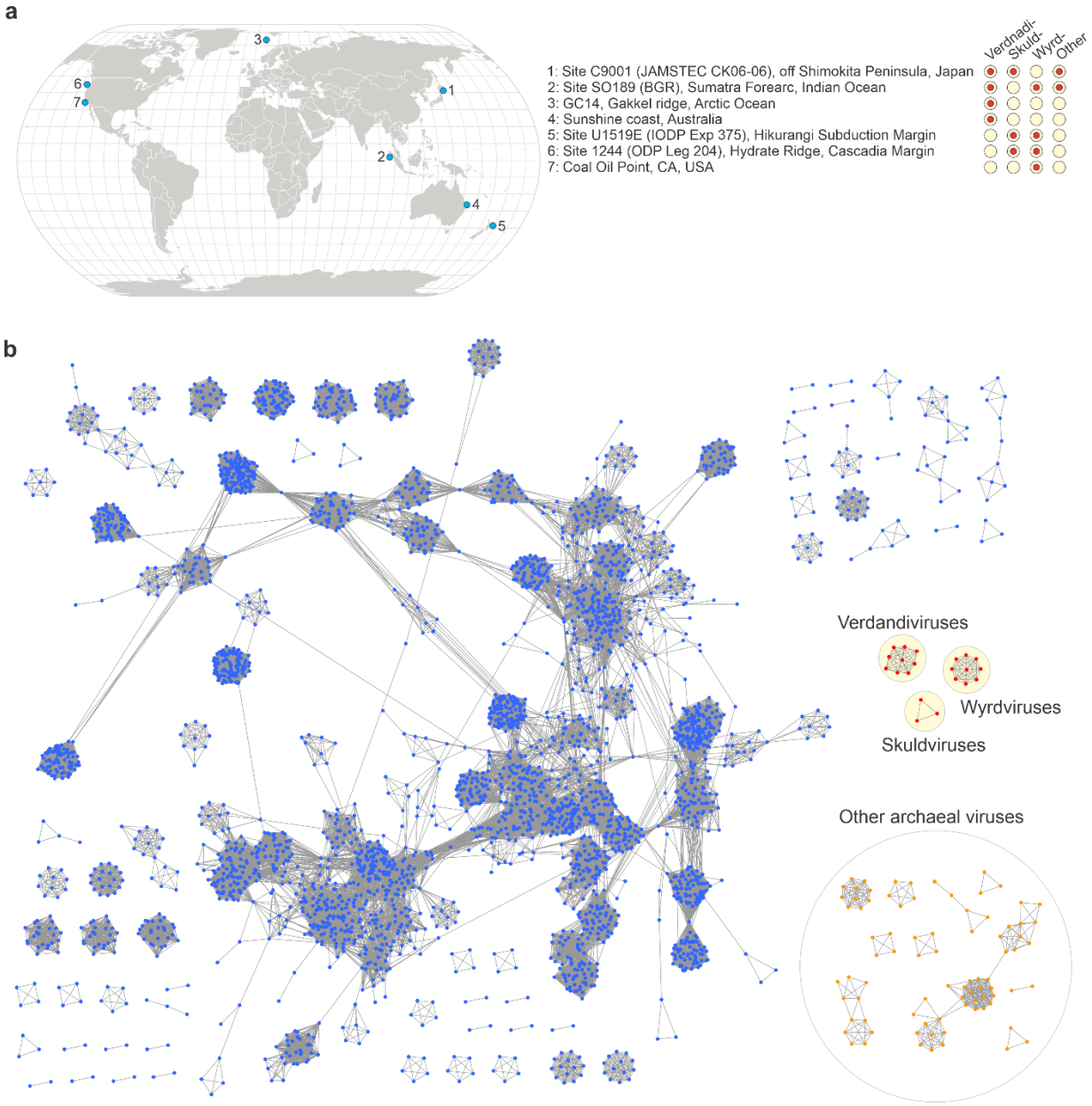
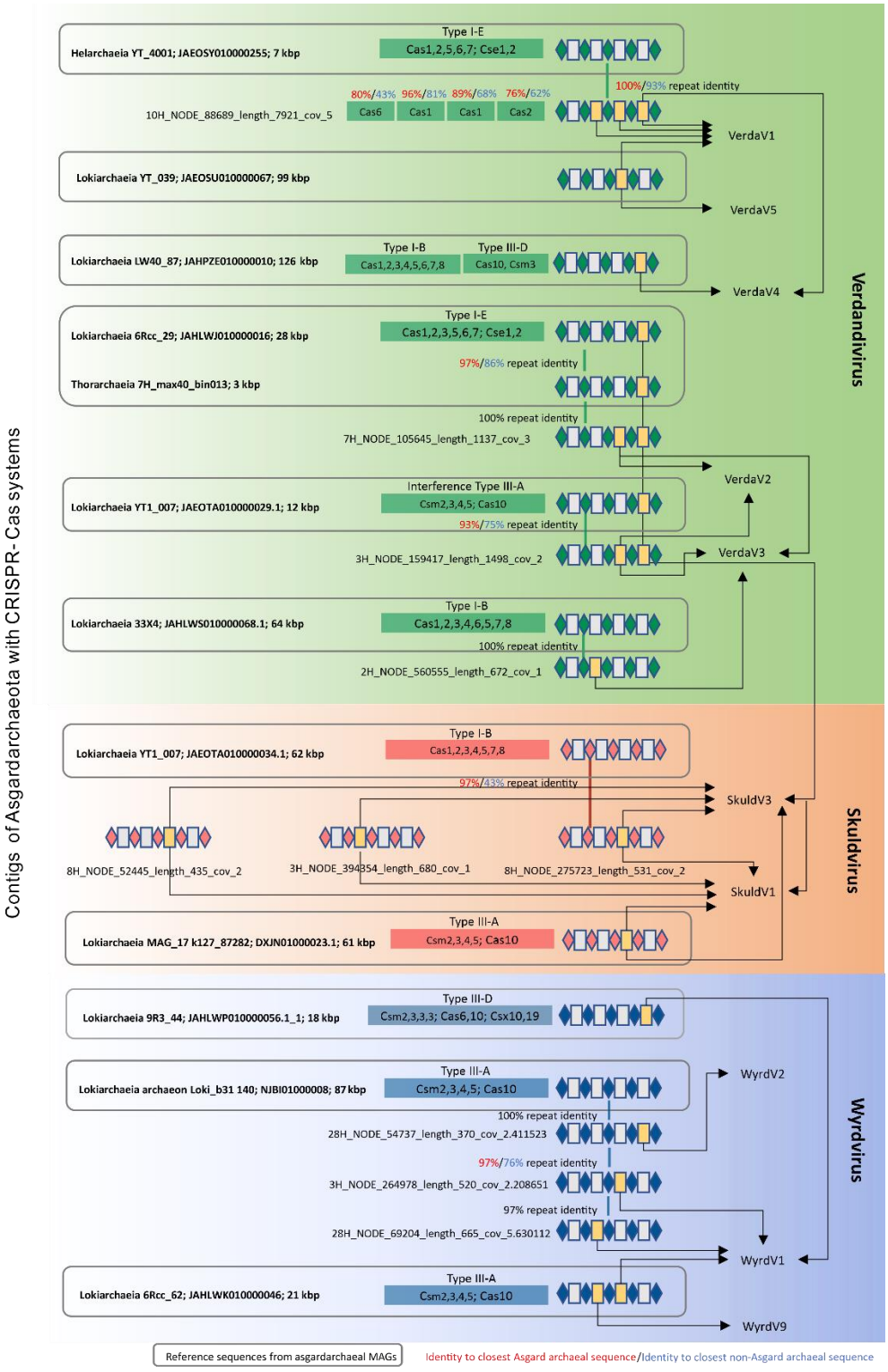
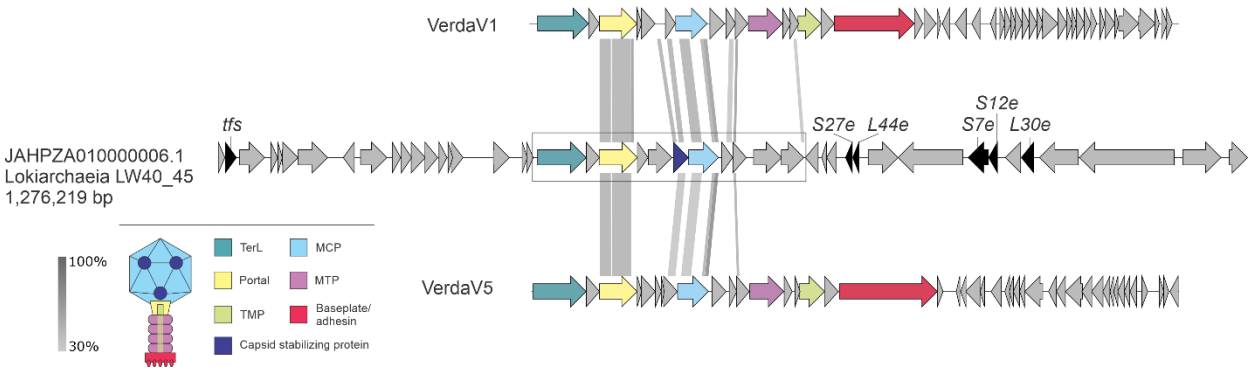


Figure 6 | Asgardarchaeal viruses and MGEs. a, Geographical distribution of asgardarchaeal viruses and MGEs. Filled circles next to geographical locations signify the presence of the corresponding virus groups in sediment samples from that location. **b,** The network-based analysis of shared protein clusters (PCs) among asgardarchaeal viruses and the prokaryotic dsDNA viruses. The nodes represent viral genomes, and the edges represent the strength of connectivity between each genome based on shared PCs. Nodes representing genomes of the three groups of asgardarchaeal viruses are in shown in red and the three groups are circled with yellow background. Nodes corresponding to other bacterial and archaeal viruses are shown in blue and orange, respectively.

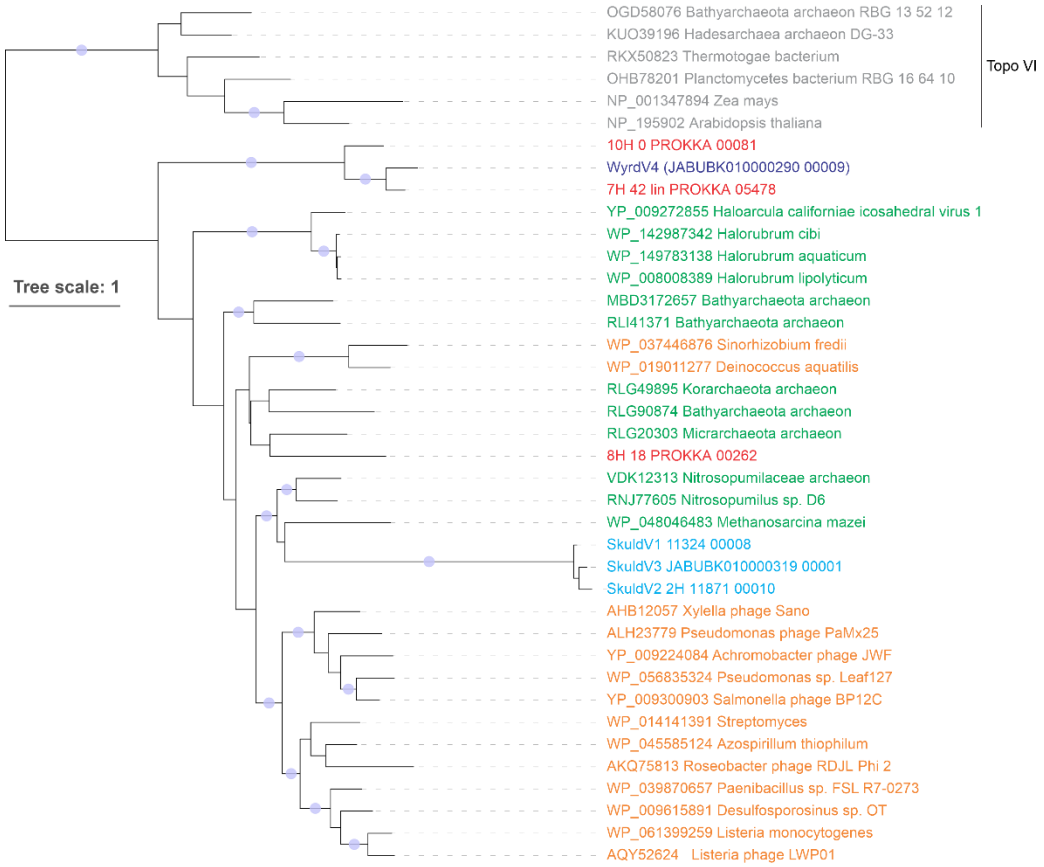
Extended Data



Extended Data Figure 1 | Matches between asgardarchaeal CRISPR spacers and viruses. The figure is divided into three blocks (green, red and blue) corresponding to the three groups of asgardarchaeal viruses, namely, verdandiviruses, skuldviruses and wyrdviruses. CRISPR repeats and spacers are indicated as diamonds and boxes, respectively. Spacers matching asgardarchaeal viruses are shown in yellow and are connected to the names of targeted viruses with arrows. Thick vertical lines connect related repeats and the exact pairwise sequence identities are indicated.



Extended Data Figure 2 | Partial provirus integrated within a genomic contig of Lokiarchaeia. The verdandivirus-derived region is boxed. Homologous genes are shown using the same colors and the key is provided on the left of the figure. Housekeeping cellular genes are shown in black and include those encoding transcription factor S (TFS) and ribosomal proteins S7e, S12e, S27e, L30e and L44e. Grey shading connects genes displaying sequence similarity at the protein level, with the percent of sequence identity depicted with different shades of grey.



Extended Data Figure 3 | Maximum likelihood phylogenetic tree of Topo mini-A proteins. Proteins of skuldviruses and wyrdviruses are shown in cyan and dark blue, respectively, whereas those encoded by other asgardarchaeal MGE are shown in red. Other Topo mini-A homologs encoded by archaea and bacteria (or their corresponding viruses) are shown in green and orange, respectively. Topo VI proteins were used as an outgroup and are colored grey. The tree was constructed using the automatic optimal model selection (LG+R5). The scale bar represents the number of substitutions per site. Circles at the nodes denote aLRT SH-like branch support values larger than 90%.

Extended Data Figure 4 | Sequence alignment of the major capsid proteins of selected wyrdviruses, fusellovirus SSV1 and halspivirus His1. TMD, transmembrane domain.

SUPPLEMENTARY INFORMATION

Supplementary text

Enigmatic MGEs of asgardarchaea

Two circular contigs, 7H_11 and 8H_18, were nearly identical and targeted by identical spacers affiliated to Thorarchaeia (Tables S3 and S4), but were assembled from metagenomes originating from samples collected at different depths (59.5 and 68.8 mbsf, respectively). Notably, 7H_42 discovered in our samples from the offshore Shimokita Peninsula, Japan was found to be related to Ga0114923_10000127 and Ga0209976_10000148 originating from the sediment samples from the Sumatra, Indian Ocean (Fig. S1), attesting to the consistency of this emerging group of Asgard MGEs.

The seven MGEs encode diverse proteins involved in DNA replication, repair and metabolism, which are common in MGE and viral genomes but, with the exception of 7H_42, Ga0114923_10000127 and Ga0209976_10000148 which form a group, display little overlap in gene content (see below). Nevertheless, 8H_18, 10H_0 and 7H_42 encode Topo mini-A homologs. Phylogenetic analysis of these proteins showed that, whereas 10H_0 and 7H_42 formed a clade with the Topo mini-A of wyrdvirus WyrdV4, 7H_42 branched among other archaeal sequences (Extended Data 4), suggestive of active exchange of Topo mini-A genes among archaeal viruses and MGEs. The 10H_0 and 7H_42-like MGEs as well as verdandiviruses and wyrdviruses encode multiple non-orthologous Zn-finger proteins, which might be involved in transcription regulation or mediate protein-protein interactions. 10H_0 and 7H_42-like MGEs also share homologs of proliferating cellular nuclear antigen (PCNA) and transcription initiation factor B (TFB) (Fig. S1), both of which have been previously identified in archaeal viruses. For instance, PCNA is encoded by several tailed archaeal viruses infecting halophilic archaea[1,2] and spindle-shaped viruses of Nitrososphaeria[3], whereas TFB homologs are encoded by certain rod-shaped viruses infecting hyperthermophilic Thermoproteota[4]. 10H_0 also encodes Cdc6/Orc1-like origin recognition protein, nucleoside 2-deoxyribosyltransferase, DNA lyase, MazG-like nucleotide pyrophosphohydrolase and bifunctional (p)ppGpp synthase/hydrolase as well as several DNA methyltransferases and nucleases. By contrast, 7H_42-like MGEs encode a DNA primase-superfamily 3 helicase fusion protein that are commonly found in diverse MGEs including diverse varidnaviruses infecting eukaryotes, a Rad51-like recombinase, several nucleases and chromatin-associated proteins containing the HMG domain. The larger elements also encode auxiliary metabolic genes, including PAPS reductase, sulfatase, methylthiotransferase, and enzymes involved in carbohydrate metabolism, which could boost the metabolic activities of the respective hosts. For the smaller contigs, 8H_18 and 8H_67, the vast majority of genes were refractory to functional annotation even using the most sensitive available sequence similarity detection tools, such as HHpred[5].

Auxiliary gene content of asgardarchaeal viruses

By dsDNA virus standards, genomes of verdandiviruses, skuldviruses and wyrdviruses are relatively small (≤20 kb). Thus, the corresponding gene contents are streamlined to include largely the core functions required for virion morphogenesis and genome replication. Nevertheless, some of these viruses encode auxiliary functions, including metabolic genes. In particular, verdandivirus VerdaV1 (and 10H_0 MGE) encode phosphoadenosine phosphosulfate (PAPS) reductase (also known as CysH), an enzyme reducing 3′–phosphoadenylylsulfate to phosphoadenosine-phosphate using thioredoxin as an electron donor. PAPS reductases have been previously identified in certain bacteriophages [6-8] and tailed haloarchaeal viruses [1], where they are thought to confer selective advantage to the host cells through facilitating sulfur metabolism and/or synthesis of sulfur-containing amino acids[7]. PAPS reductase of VerdaV1 might perform a similar function.

Wyrdviruses WyrdV2 and WyrdV6 carry a block of three genes coding for dUTPase, thymidylate synthase X (ThyX) and an uncharacterized protein that is conserved in some phages and is annotated as nucleotide modification associated domain 1 protein (PF07659.13, DUF1599) (Fig. 5). This putative operon is likely to be involved in the biosynthesis of thymidylate from dUTP, to increase the pool of nucleotides available for the synthesis of viral DNA. WyrdV3 encodes a homolog of the nucleoside pyrophosphohydrolase MazG, which in bacteria prevents programmed cell death by degrading the central alarmone, ppGpp[9]. MazG is highly conserved in tailed bacteriophages infecting cyanobacteria[10]. Biochemical characterization of a cyanophage MazG has shown that, instead of degrading ppGpp, it preferentially hydrolyses dGTP and dCTP[11]. Thus, MazG homolog in WyrdV3 might either function in disarming antiviral systems triggered by nucleotide-based alarmones, such as ppGpp, or in adjusting the intracellular nucleotide concentrations for optimal viral genome synthesis. Notably,

MazG homologs are also encoded by asgardarchaeal MGEs 10H_0, 7H_42, Ga0114923_10000127 and Ga0209976_10000148 (Fig. S2).

None of the known archaeal viruses encodes its own RNA polymerase[12]. Nevertheless, various transcription regulators with HTH, Zn-finger or ribbon-helix-helix domains are abundantly encoded in archaeal virus genomes[13]. This is also the case with asgardarchaeal viruses described herein. Verdandiviruses and wyrdviruses encode multiple non-orthologous Zn-finger proteins, whereas skuldviruses encode several proteins with HTH domains (Fig. 4a). In addition, WyrdV1 (as well as 10H_0, 7H_42, Ga0114923_10000127 and Ga0209976_10000148) encodes a transcription initiation factor B (TFB), a homolog of eukaryotic TFIIB, which guides the initiation of RNA transcription[14]. Among archaeal viruses, TFB homologs have been previously identified only in certain rod-shaped viruses infecting hyperthermophilic archaea [4]. Thus, Asgard viruses appear to fully rely on the core transcription machinery of their hosts but encode various transcription factors that could be involved in the recruitment of this machinery for expression of viral genes as well as in the regulation of virus gene transcription. As mentioned above, some of the genes regulated by these transcription factors are likely to encode antidefense proteins.

WyrdV1 and WyrdV3 encode homologs of the carbohydrate-specific 3'-O-methyltransferase[15]. In many archaeal viruses, the structural proteins are glycosylated by either the virus or host encoded glycosyltransferases, although the biological role of this post-translational modification remains unclear. The methyltransferase of WyrdV1 and WyrdV3 could participate in modification of the glycans attached to the virion proteins.

Supplementary references

1 Mizuno, C. M. *et al.* Novel haloarchaeal viruses from Lake Retba infecting Haloferax and Halorubrum species. *Environ Microbiol* **21**, 2129-2147, doi:10.1111/1462-2920.14604 (2019).

2 Raymann, K., Forterre, P., Brochier-Armanet, C. & Gribaldo, S. Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol Evol* **6**, 192-212, doi:10.1093/gbe/evu004 (2014).

3 Kim, J. G. *et al.* Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. *Proc Natl Acad Sci U S A* **116**, 15645-15650, doi:10.1073/pnas.1905682116 (2019).

4 Baquero, D. P. *et al.* New virus isolates from Italian hydrothermal environments underscore the biogeographic pattern in archaeal virus communities. *ISME J* **14**, 1821-1833, doi:10.1038/s41396-020-0653-z (2020).

5 Gabler, F. *et al.* Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics* **72**, e108, doi:10.1002/cpbi.108 (2020).

6 Mara, P. *et al.* Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *ISME J* **14**, 3079-3092, doi:10.1038/s41396-020-00739-3 (2020).

7 Summer, E. J., Gill, J. J., Upton, C., Gonzalez, C. F. & Young, R. Role of phages in the pathogenesis of Burkholderia, or 'Where are the toxin genes in Burkholderia phages?'. *Curr Opin Microbiol* **10**, 410-417, doi:10.1016/j.mib.2007.05.016 (2007).

8 Farlow, J. *et al.* Genomic characterization of three novel Basilisk-like phages infecting Bacillus anthracis. *BMC Genomics* **19**, 685, doi:10.1186/s12864-018-5056-4 (2018).

9 Gross, M., Marianovsky, I. & Glaser, G. MazG -- a regulator of programmed cell death in Escherichia coli. *Mol Microbiol* **59**, 590-601, doi:10.1111/j.1365-2958.2005.04956.x (2006).

10 Sullivan, M. B. *et al.* Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ Microbiol* **12**, 3035-3056, doi:10.1111/j.1462-2920.2010.02280.x (2010).

11 Rihtman, B. *et al.* Cyanophage MazG is a pyrophosphohydrolase but unable to hydrolyse magic spot nucleotides. *Environ Microbiol Rep* **11**, 448-455, doi:10.1111/1758-2229.12741 (2019).

12 Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D. & Koonin, E. V. Viruses of archaea: Structural, functional, environmental and evolutionary genomics. *Virus Res* **244**, 181-193, doi:10.1016/j.virusres.2017.11.025 (2018).

13 Iranzo, J., Koonin, E. V., Prangishvili, D. & Krupovic, M. Bipartite network analysis of the archaeal virosphere: Evolutionary connections between viruses and capsidless mobile elements. *J Virol* **90**, 11043-11055, doi:10.1128/JVI.01622-16 (2016).

14 Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol* **9**, 85-98, doi:10.1038/nrmicro2507 (2011).

15 Bernard, S. M. *et al.* Structural basis of substrate specificity and regiochemistry in the MycF/TylF family of sugar O-methyltransferases. *ACS Chem Biol* **10**, 1340-1351, doi:10.1021/cb5009348 (2015).

Supplementary figures

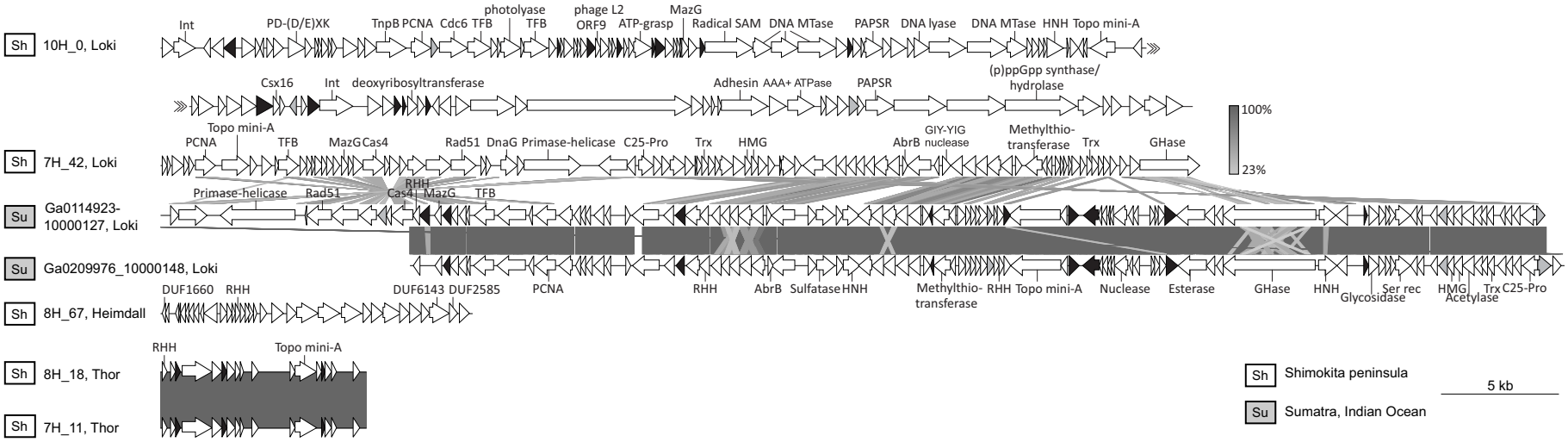
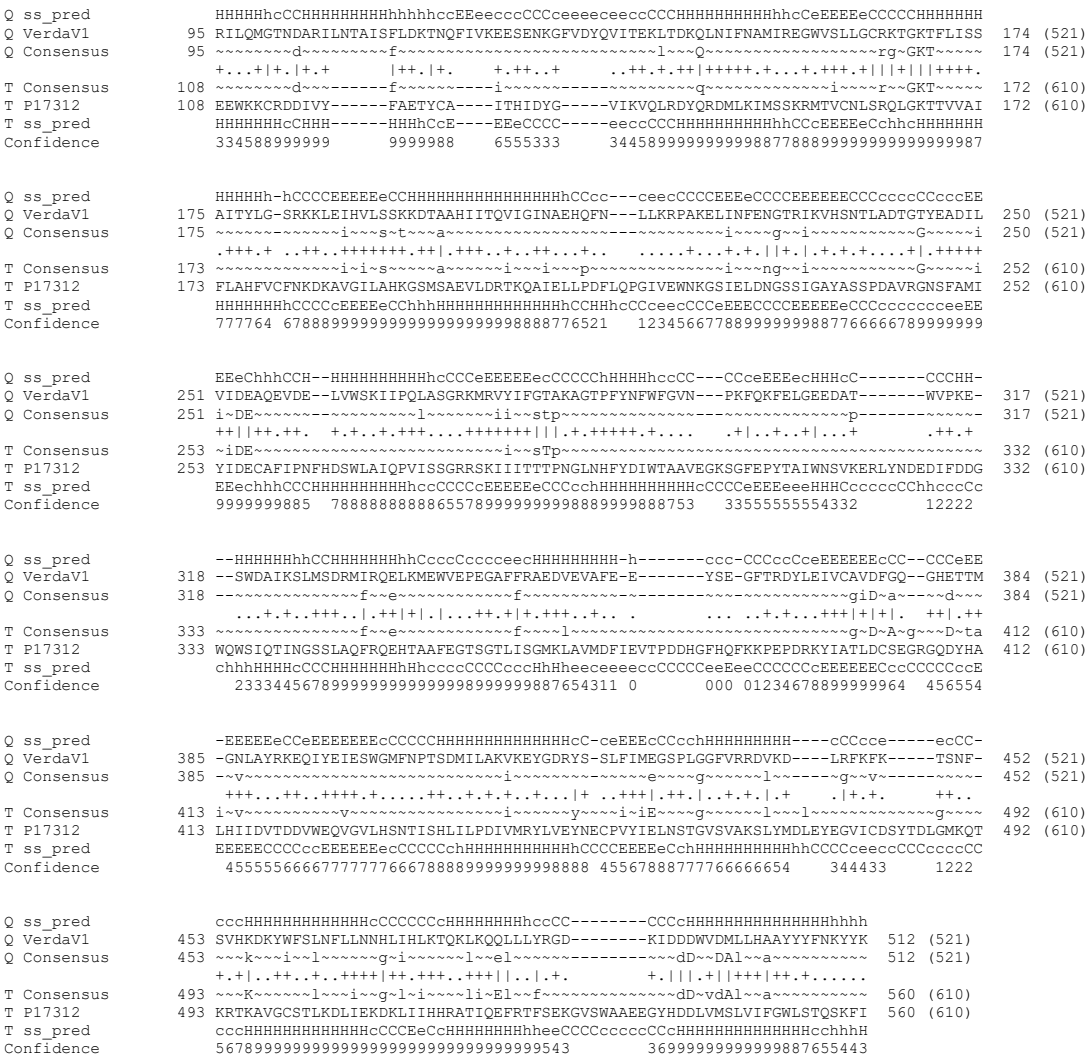


Figure S1. Genome maps of asgardarchaeal CRISPR-targeted MGEs. Due to the lack of known virus hallmark genes, the MGEs are not identified as viruses, but could represent novel virus groups. Genes encoding putative DNA-binding proteins with Zn-binding and helix-turn-helix domains are colored in black and grey, respectively. Grey shading connects genes displaying sequence similarity at the protein level, with the percent of sequence identity depicted with different shades of grey Abbreviations: Int, integrase; PD-(D/E)XK, PD-(D/E)XK family nuclease; PCNA, proliferating cellular nuclear antigen; TFB, transcription initiation factor B; MTase, methyltransferase; HNH, HNH family nuclease; PAPSR, phosphoadenosine phosphosulfate reductase; C25-Pro, C25-family protease; Trx, thioredoxin; GHase, glycoside hydrolase; RHH, ribbon-helix-helix domain-containing DNA binding protein; Ser rec, serine superfamily recombinase; HMG, high mobility group domain-containing chromatin-associated protein.

a

>P17312 TERL_BPT4 Terminase, large subunit OS=Enterobacteria phage T4 OX=10665 GN=17 PE=1 SV=1
Probab=100.00 E-value=4.1e-34 Score=301.28 Aligned_cols=403 Identities=15% Similarity=0.129 Sum_probs=304.0 Template_Neff=10.300



b

>A0A0U5AF03 CAPSD_BPK22 Major capsid protein OS=Pseudomonas phage KPP22 OX=1772250 PE=1 SV=1
Probab=99.97 E-value=3.5e-29 Score=244.70 Aligned_cols=275 Identities=11% Similarity=0.022 Sum_probs=0.0 Template_Neff=7.800

```
Q ss_pred              CcccCCCHHHHHHHHHHHHhhcchhccccCCCCCCCC-ccceEEEeeeeeeceecc--ccCCcccccccc-cceEEEEe
Q VerdaV1             2 VEIHGVSERQWKGIDTQVAQLMREPIVYTRLPSGMQIG-KGKNKAEYFQATDIFYVN--EGNRFSYEDKNT-GARKIRDA   77 (318)
Q Consensus           2 ~~~~~l~~~~w~~Id~~v~~~~~~~~~L~ar~l~p~v~~~~-G~~~~~~~~~~~~~d~~~~~~--~g~~~~~~~~v~~-~~~~~~~~~v   77 (318)
                        ...++++.++|.+||.+|++.++..+.+|+|++ +...  .|+.....++.++..+.+  +++......+.+ ......++
T Consensus        73 ~~~~~ip~~~l~~id~~vi~~~~~~~~~~~li~-v~~~g~w~~~~~~~~~~~~~e~~G~a~~~~d~~d~p~~~~~~~~~~~~~~~v  151 (382)
T A0A0U5AF03       73 TPSIPTPIQFLQTWLPGLVKVMTAARKIDEIIG-IDTVGSWEDQEIVQGIVEPAGTAVEYGDHTNIPLTSWNANFERRTI  151 (382)
T ss_pred              CCCCchHHHHHHHhcchhcccccccccHHHHcC-cccCCCceeEEEEEEEecCCCceEEEecCCCCCCceeeeEeeceeeee


Q ss_pred              EEEEEEecccccHHHHHHHHhcCCCCchHHHHHHHHHHHHHHHhCeEeecCC---ccCcccceecccCCCcccCCCceeee
Q VerdaV1            78 IITYLTLPIHIDEKDIDASSSEGITPLLVQMEVECRAKMQEKIQKIILHGSA---HTGGRGICTFQQDGDVVEQTYAFEA  154 (318)
Q Consensus          78 ~~i~~~f~i~~~~rel~as~~~g~~~ld~~~a~~A~r~vae~ed~lif~G~~---~~gi~GL~n~~~~~~~~~~~g~~~~~~  154 (318)
                        +.++..|.++  ++|++++++.|. +++.+++.+|++.++++++|+++|+|++    ..+++||+|+        ++.+...
T Consensus       152 ~~~~~~~~~~~~---~~El~~a~~~g~~~l~~~k~~aa~~~~~~~~~n~i~f~G~~~~~~~~~~~~~GLlN~--------p~l~~~~  220 (382)
T A0A0U5AF03      152 VRGELGMMVG--TLEEGRASAIRL-NSAETKRQQAAIGLEIFRNAIGFYGWQSGLGNRTYGFLND--------PNLPPFQ  220 (382)
T ss_pred              EEEEEEeEec--hHHHHHHHhcCC-CHHHHHHHHHHHHHHHHHHhhEeeeecCCCCCceeecccC--------CCCCccC


Q ss_pred              c---Cccc--chHHHHHHHHHHHHHHH--HHCCCCCC----CEEEEChhHHHHHhcccCCCCCCcHHHHHHHHHHhcCCcc
Q VerdaV1           155 T---LTWT--NNLQFSLEMEIARGLL--RAAHIMPP----YKLWMTPGIRPTLRANVNATTGVSDWKLFQQDWMNSGVIS  223 (318)
Q Consensus         155 ~----~~W~--t~~~i~~dv~~ai~~L----~~g~~~p-----~~L~L~p~~y~~L~~~~~~~~~~t~~~~~~~~~i~~~i~~~~~~~~~~  223 (318)
                        .    .+|.  ++.++++||..++.+|  .+.++++|    ++|+||++.|..|.+  .+..+.    +++++|+++.
T Consensus       221 ~~~~~~~W~~~T~~eI~~dv~~~~~~~l~~~s~g~~~p~~~~~~tL~lp~~~~~~~L~~~~~~~~~~~~~~~~~~tvl~~Lk~~~~~---  291 (382)
T A0A0U5AF03      221 TPPSQGWSTADWAGIIGDIREAVRQLRIQSQDQIDPKAEKITLALATSKVDYLSV--TTPYGI----SVSDWIEQTY---  291 (382)
T ss_pred              CCCCCCccccCCHHHHHHHHHHHHHHHHHHHHcCCCCCCCCCCCCCceEEEEeHHHHHHhcc--CCCCCC----cHHHHHHHHC---


Q ss_pred              CCCCCCCCCCchhhceecchh---cCCCCCCCCceEEEEEEECCC-----------CeEEEEeccCcEecCCCCCCCceeece
Q VerdaV1           224 GEKGTEPFIPDINQITETDAL---YNGTLSTATQAFLLWKDDI-----------NCNYIAEAYPIHRAGIPNKEFEGDID  289 (318)
Q Consensus         224 ~~~~~~~~~~~~~~~~~~~i~~~~~~~~~~~~~~~~~~~~~~~~~~~d~~l~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~i~~~~~~~~~~~~~~~~~~~~~~~~~~~~~i~  289 (318)
                              |.+. |..++++++.   +.+    +.++++++.+++           +++.+.+|++++.++.+.+++.|.++
T Consensus       292 ------p~i~~-I~~~pel~~a~~~g~~------g~~~~~l~~~~~~~~~~~~~~~g~~~~~~~~~~~~~~~~vp~~~~~~~~~~~~~~~~~~~~~~~~~  360 (382)
T A0A0U5AF03      292 ------PKMR-IVSAPELSGVQMKAQE----PEDALVLFVEDVNAAVDGSTDGGSVFSQLVQSKFITLGVEKRAKSYVED  360 (382)
T ss_pred              ------CCCE-EEECHhhhcccccCCC----CccEEEEEEECchhcccCCCCCcccceeeeeccccceecccEeccceEEEe


Q ss_pred              EEEeeecEEeecCCCceEEE
Q VerdaV1           290 YALRWAGCFLPKNPKGAVYV  309 (318)
Q Consensus         290 ~~~~r~gGv~~~~~~~~~~~~~~~  309 (318)
                        |..+++|+++ |.|..|++.
T Consensus       361 ~~~~t~G~~i~~rP~ai~~~  379 (382)
T A0A0U5AF03      361 FSNGTAGALC-KRPWAVVRY  379 (382)
T ss_pred              eeceeeeEEE-eccceehhh
```

Figure S2. Results of the HHsearch analysis queried with the putative large terminase subunit (a) and the major capsid proteins (b) of verdandivirus VerdaV1. H(h), α-helix; E(e), β-strand; C(c), coil.

a

```
>P15794 CAPSD_BPPM2 Major capsid protein P2 OS=Pseudoalteromonas phage PM2 OX=10661 GN=II PE=1 SV=2
Probab=98.32  E-value=0.0012  Score=61.40  Aligned_cols=219  Identities=12%  Similarity=0.056  Sum_probs=0.0  Template_Neff=8.200

Q ss_pred              eeecCceeeecceeeccEEEEEecCCccccCcchhheeeccccceEEEEECCEEEEEE-eccccCccccchHHHHHHHHHHH
Q SkuldV1            15 AESENKHVKIDNTFPIKKIVIIHNPGDFAGGSTGVLIANTILDNLRVFVHGKEVISL-IGDVDVDTAPFAIQLMREANKL   93 (277)
Q Consensus          15 a~~~~~~~~~ti~~tf~ik~I~I~~~~~ga~sGGatG~~VaN~a~~s~rv~~nGk~iI~~-dG~~~vd~~s~gi~lLRE~~~~~   93 (277)
                        +.+.+.+++|+.-.....|.+..+            .+...-++.|||.+||+.++.+ +|              +.|..+|++
T Consensus        14 ~~g~~at~~lp~g~tY~~i~l~~t------------~~t~a~I~~I~v~~NG~~i~~~~~~~~~~~~~~~~~~~~-----------~~L~~~n~y   71 (269)
T P15794           14 AAGNSCSIKLPIGQTYEVIDLRYS-----------GVTPSQIKNVRVELDGRLLSTYKTL-----------NDLILENTR   71 (269)
T ss_pred              CCCCEEEEEcCCCCCeeEEEEEEEEEc-----------CCCHHHcCeEEEEEECCEEEEEEeCCH-----------HHHHHHHHH


Q ss_pred              hcccCCccceeEEEcCCccCCC-------------------CeEEEEEEecceeeecCCCCCceeeeceEEEEEEEeCCCCe
Q SkuldV1            94 RNKVADADEYFEIDFPKAIGRG-------------------HNAYIDFRFNTIANMNDGDRTTYTGATIDIMVEVGTPGR  154 (277)
Q Consensus          94 ~~~VA~~d~y~~I~FP~aip~g-------------------~dvqI~i~~~tsaq~~gGDr~T~ta~t~DI~v~~~~kG~  154 (277)
                        . +-+..++++.|.|..+--++                    .+.+|++.+++.+.          +-+++.......+..
T Consensus        72 ~-g~~~~g~l~i~F~~~~~~~~~~~~~~~~~~~~~~~~~~~~al~t~~~~s~~lev~i~a~~~~~~~~~~~~~~~~~~~~p~L~~~a~~~~~~~~~~  140 (269)
T P15794           72 H-KRKIKAGVVSFHFVRPEMKGVNVTDLVQQRMFALGTVGLTTCEIKFDIDEAAA----------GPKLSAIAQKSVGTA  140 (269)
T ss_pred              h-cCCccCcEEEEecCChhhcCCCchhHheehhcccCCCccEEEEEEEeCCCCC----------CCEEEEEEEeCCCCC


Q ss_pred              EEEeeEcceeeccccccCCCccccCCcc--ccceeeeEEEEEecCCCcccccceeeEEEeeCceeeecC-cHHHHHHHHccC
Q SkuldV1           155 LSLLRNCGKIVLGANTGRLPEFINPT--EFGYKALNLMVLIEDDGTPSNTAIAKLELRKGSNIIREG-SIPKLQELTKGK  231 (277)
Q Consensus         155 ~ip~~~~g~~s~ga~~G~l~e~lp~~~--~~plR~~~Lm~it~D~gT~S~tay~~L~Is~g~~~i~DG-siakl~~~~~~~K  231 (277)
                        .-.++....+....-.|.....--|.  ...+|.  +.+-+.|          .++++|+.++..+||. +-+.....+++-
T Consensus       141 ~~~~~~~~~~~~~~~~~~~G~~~~~~~~lp~~~~~~~~i~~--ih~~~~~~~~~~~~~~~~I~~v~v~~D~~~i~~~~~~~~~~~~~~~~~~~~~~~  210 (269)
T P15794          141 PSWLTMRRNFFKQLNNGTTEIADLPRPVGYRIAA--IHIKAAG--------VDAVEFQIDGTKWRDLLKKADNDYILEQY  210 (269)
T ss_pred              CeeEEEEeeeeeccCCCcccccccCCCCcceeEE--EEEecCC--------cCEEEEEECCEEEEecccHHHHHHHHHHh


Q ss_pred              ceeeccceEEEEEecC--cC--------CHHHeEEEEEeccCccceEEEEEeeccC
Q SkuldV1           232 FSIALGTGFVWIPIKQ--AI--------SASQLKLISQIDSAGTNIEVHWMLTSLK  277 (277)
Q Consensus         232 s~vA~s~G~~~i~fp~--ki--------s~~tLkl~~~~~~tAgt~~e~h~~~~~~~~  277 (277)
                        .++. -+|++.+.|=.  ..        ..++|+|++..++|++=--+.-.+..++
T Consensus       211 ~~~p-qag~~~~~Df~~~g~~~~~~~~~~~~~~~~~~l~~~~~~~~~~~a~~i~i~~E~l~~~~~  265 (269)
T P15794          211 GKAV-LDNTYTIDFMLEGDVYQSVLLDQMIQDLRLKIDSTMDEQAEIIVEYMGVWS  265 (269)
T ss_pred              Cccc-CCCeEEEeCCccCChhhhhhhhhccccceEEEEEEecCccceEEEEEEeccc
```

b

>3J31_B Major capsid protein; Double jelly roll fold; 4.5A; Sulfolobus turreted icosahedral virus
Probab=97.84 E-value=0.013 Score=58.95 Aligned_cols=242 Identities=13% Similarity=0.060 Sum_probs=0.0 Template_Neff=6.300

```
Q ss_pred            eeEeeecCceeeeeceeeeeEEEEEecCCcccCcchhhee-----eccccceEEEEECC-EEEEEEecccCccccchHHH
Q SkuldV1         12 IAFAESENKHVKIDNTFPIKKIVIIHNPGDFAGGSTGVLI-----ANTILDNLRVFVHG-KEVISLIGDVDVDTAPFAIQ   85 (277)
Q Consensus       12 iafa~~~~~~~ti~~tf~ik~I~I~~~~ga~sGGatG~~V-----aN~a~~s~rv~~nG-k~iI~~dG~~~vd~~s~gi~   85 (277)
                     .+++.+.++++.|+|.-+|++|.|.|.......+++++...+     ...+++.|+|..|| .-+.+++|        ...+
T Consensus       13 ~s~~a~~t~tIdLP~~~pIS~I~l~~~~t~~~s~~~~~~~~~~~p~~~I~kIelv~~Gs~vi~SlsG--------~~l~   84 (345)
T 3J31_B          13 YAWTAGTNIPIKIPRNNFIRKIRVQLIGSISNSGTAAVTLPSAPFPYNLVQTFNLSYEGSKTLYSVSG--------TGLG   84 (345)
T ss_dssp            EECCTTCEEEEEECCSSEEEEEEEEEEEEEECCSSSCEECCCTTTTGGGSCEECEEETTTEESEEEEH--------HHHH
T ss_pred            cccCCCCeEEEEcccCcceeEEEEEEEEeeeCCCCccccCCCCCchHHhceeEEEEECCCceEEEEeH--------HHHH


Q ss_pred            HHHHHHHHhcccCC----------ccceeEEEcC------CccCCCCeEEEEEEecceeeecCCCCceee-eceEEEEEE
Q SkuldV1         86 LMREANKLRNKVAD----------ADEYFEIDFP------KAIGRGHNAYIDFRFNTIANMNDGDRTTYT-GATIDIMVE  148 (277)
Q Consensus       86 lLRE~~~~~~~VA~----------~d~y~~I~FP-------aip~g~dvqI~i~~~tsaq~~gGDr~T~t-a~t~DI~v~  148 (277)
                     .|-.+++.....          .+..+.+.||      .+..++ +.|++|.++.     .++..++. +.++.++-.
T Consensus       85 aLn~y~~~~~p~~~~~~~~~~~~~~~~~~~IdFG~~~A~d~~~nLqL~It~~~~~~~a~~s~~~~~itv~~~  158 (345)
T 3J31_B          85 ILMYYTTKGQNPAYPAPGTSVPASGSVNLNVMWEFDLARFPATMVQ-NIILSILTGQ-----APSGVSINASFYITITYE  158 (345)
T ss_dssp            HHHHHHTTTCSCEESCTTCEECTTSEEEEEEEEEEEEEEECTTSSS-CEEEEEECCC-----CCTTSEECCEECCEEEEE
T ss_pred            HHHHHHhCCCCCCCCCCcCCCCCceeEEEEEEccceeccCCC-CeEEEEECC-----CCCCCccCceeEEEEEE


Q ss_pred            ----------eCCCCe-------EEEeeEcce-eeccccccCCCcccCCccccccceeeeEEEEEecCC--Ccccceeee--
Q SkuldV1        149 ----------VGTPGR-------LSLLRNCGK-IVLGANTGRLPEFINPTEFGYKALNLMVLIEDDG--TPSNTAIAK--  206 (277)
Q Consensus      149 ----------~~~kG~~~~~~~~~ip~~~~g~~~s~ga~~G~l~e~lp~~~~plR~~~Lm~it~D~g--~T~S~tay~~~~  206 (277)
                                ++.++.       ..+-...-. ++-.+..+.+.-...-=|...|||+  +++..+|++  -.+++..+.
T Consensus      159 ~~~~~~~~~~~~~~~~mp~~~~~kev~~~~~~~~~s~~~~idLPtg~~~Rk--lli~~d~~~g~~~~~~~~~l  236 (345)
T 3J31_B         159 RVTAQEILSEGGLGADGEMPLATVLPKVIEIPTFNVPASSAPIHVAYLQPGQIYKR--QLVYVINSTSGINNTDPTEYEL  236 (345)
T ss_dssp            CCCHHHHHTTTSBCSSCSSBTTCCEECEEEEEEEEEECBCSSCEEEEECCSSSEEEE--EEEEEEBTTTBTCCCCEEEEEE
T ss_pred            EecccceeecCCCCcccCCCchhccceEEEeecCCCCCCCCceEEEECCCCCeEeE--EEEEEeCCCCCCCCCcccccce


Q ss_pred            -EEEeeCceeeeecCcHHHHHHHHccCce-eeccceEEEEEecC----------c--CCHHHeEEEEEeccCccceEEE
Q SkuldV1        207 -LELRKGSNIIREGSIPKLQELTKGKFS-IALGTGFVWIPIKQ----------A--ISASQLKLISQIDSAGTNIEVH  270 (277)
Q Consensus      207 -L~Is~g~~~i~DGsiakl~~~~~~Ks~~vA~s~G~~~i~fp~----------k--is~~tLkl~~~~tAgt~~e~h  270 (277)
                      |....+. ..++-...-|+++.+..-+ -.+..|++++-|-+          .  -..++++|.+..+.+|.---||
T Consensus      237 ~l~~~~~~~~~~~~~~l~~~~~~~~~~~~~g~~~iDf~~~~g~~l~l~~~~~~~~~L~~~~~~g~~~vi~  312 (345)
T 3J31_B         237 KIVRGVPT-DKIKVSWAALQAENQAEYQVAPYSGASAIIDFRKYFNGDLDLTHAPSDSIEYDLALQNQDNVYSLYVS  312 (345)
T ss_dssp            EECSSSCE-EEEECBHHHHHHHHHHHHCSCCSSTTEEEEEGGGTTTSSEECSSSCSSSCEEEEEESSCEEEEEEEEE
T ss_pred            EEeecCCC-ceeeeCHHHHHHHHHhhhCCCCCCCcEEEEechhhccCccccCCCCCCCcEEEEEEecCCCceeEEEE


Q ss_pred            HHHHHHHHHHhcccCc-----------------------------cceeEEEcCCccCCCC------------eEEEE
Q SkuldV1         84 IQLMREANKLRNKVADA-----------------------------DEYFEIDFPKAIGRGH------------NAYID  121 (277)
Q Consensus       84 i~lLRE~~~~~~VA~~-----------------------------------d~y~~I~FP~aip~g~------------dvqI~  121 (277)
                        |+.++++++.....                                 .-.|.+++|-++-++.           +++++
T Consensus      117 ---L~~~n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~as~~Ag~t~~v~f~l~iPval~~~D~~G~il~qn~~t~l~L~  193 (395)
T P22535         117 ---LHFVNTAKQGAPFLSSMVTDSPIKYGDVMNVIDAPATIAAGATGELTMYYWVPLAYSETDLTGAVLANVPQSKQRLK  193 (395)
T ss_pred            ---HHHHHHHhcCCCeeeeccCCCCcccCCcccccCcceeccCCCeEEEEEEeeeeecCCccceeeeeCCCCccEEEE


Q ss_pred            EEecc---------------eeeecCCCCceeeeceEEEEEEe-----CC-CCeEEE---------eeEccceeeccccC
Q SkuldV1        122 FRFNT---------------IANMNDGDRTTYTGATIDIMVEV-----GT-PGRLSL---------LRNCGKIVLGANTG  171 (277)
Q Consensus      122 i~~~t---------------saq~~gGDr~T~ta~t~DI~v~~~~~~~~kG~~ip-------------g~~s~ga~~G  171 (277)
                     |.+++               .-+..+++..|+..=.+.-+.|+     -. .+...|         +|..-...-.+..-
T Consensus      194 Vt~at~~~~~~~~~~~~a~~y~~~~~t~tl~tv~V~~~~e~~~iPi~~~~~~P~~dlst~~~l~~~~~~~~~G~  273 (395)
T P22535         194 LEFANNNTAFAAVGANPLEAIYQGAGAADCFEEISYTVYQSYLDQLPVGQNGYILPLIDLSTLYNLENSAQAGLTPNVD  273 (395)
T ss_pred            EEEcCccccccccCCChhHHhcCCCcceEEeEEEEEEEEeEeecCCCCCCCccCCCccHHHHHhHhheeeccCCCCce


Q ss_pred            CCcccCCccccceeeeEEEEEecCCCccc--cceeeEE-EeeCceeeeecCcHHHHHHHccCceeeeccceEEEEEecC-
Q SkuldV1        172 RLPEFINPTEFGYKALNLMVLIEDDGTPSN--TAIAKLE-LRKGSNIIREGSIPKLQELTKGKFSIALGTGFVWIPIKQ-  247 (277)
Q Consensus      172 ~l~e~lp~~~~plR~~~Lm~it~D~gT~S~--tay~~L~~Is~g~~~i~DGsiakl~~~~~Ks~vA~s~G~~~i~fp~~  247 (277)
                     +.-++  |...-||.  |++ ++|.|.|.+  +..+.+. +..t...+++-...-+++.++.+.+-.+-.|++++.|-.
T Consensus      274 ~~i~L~~p~g~~yr~~~l~~~v~DnG~~~~~~~d~~~~~l~~n~~di~~~~~~~~~~~~~~~~~~~lp~Gvy~~D~~~  348 (395)
T P22535         274 FVVQY--ANLYRYLS--TIA-VFDNGGSFNAGTDINYLSQRTANFSDTRKLDPKTWAAQTRRRIATDFPKGVYYCDNRDK  348 (395)
T ss_pred            eEEEc--ccccceEE--EEE-EEeCCCcCCCCccccccceeeeCCceeeccCHHHHHHHHHHhCCCCCeEEEEeCCCC


Q ss_pred            ---cCCHHHeEEEEEeccCccceEEEEEee
Q SkuldV1        248 ---AISASQLKLISQIDSAGTNIEVHWMLT  274 (277)
Q Consensus      248 ---kis~~tLkl~~~~tAgt~~e~h~~~~  274 (277)
                        ----+.+.|.+..-+.|.+-.+-+-|.
T Consensus      349 gi~t~~~~~~~l~~~~~~~~~~~~~~~~~  378 (395)
T P22535         349 PIYTLQYGNVGFVVNPKTVNQNARLLMGYE  378 (395)
T ss_pred            ccccccccceEEEEccccccccceeEEEEe
```

c

```
>P22535 CAPSD_BPPRD Major capsid protein P3 OS=Enterobacteria phage PRD1 OX=10658 GN=III PE=1 SV=2
Probab=96.16  E-value=1.5  Score=46.49  Aligned_cols=247  Identities=11%  Similarity=0.091  Sum_probs=0.0  Template_Neff=5.600

Q ss_pred            eeEeee--cCceeeeccceeeccEEEEEecCCcccCcchhheeecc-----ccceEEE-EECCEEEEEEecccCccccchH
Q SkuldV1         12 IAFAES--ENKHVKIDNTFPIKKIVIIHNPGDFAGGSTGVLIANT-----ILDNLRV-FVHGKEVISLIGDVDVDTAPFA   83 (277)
Q Consensus       12 iafa~~~~~~~~ti~~tf~ik~I~I~~~~ga~sGGatG~~VaN~------a~~s~rv~~nGk~iI~~dG~~~vd~~s~g   83 (277)
                     +.|...  ...+++|+|+--+++|.+.+.+.....|++|+.-...     +++.||+ -.|+-..++++|..
T Consensus       45 ~~~~~~~~s~~t~~Lp~~G~L~~l~l~v~~tlt~~~~t~a~t~~~~~P~~lI~~I~l~~~gn~~~~~~sG~~~~~~~~  116 (395)
T P22535          45 QTFDPANRSVFDVTPANVGIVKGFLVKVTAAITNNHATEAVALTDFGPANLVQRVIYYDPDNQRHTETSGWH--------  116 (395)
T ss_pred            eeccCcCCcEEEEecCCCceEeeEEEEEEEEEEcCCCcccccCCcChhHhcceEEEECCCCceeEEccHHH--------

Q ss_pred            HHHHHHHHHhcccCCc-----------------------------cceeEEEcCCccCCCC------------eEEEE
Q SkuldV1         84 IQLMREANKLRNKVADA-----------------------------DEYFEIDFPKAIGRGH------------NAYID  121 (277)
Q Consensus       84 i~lLRE~~~~~~VA~~-----------------------------------d~y~~I~FP~aip~g~------------dvqI~  121 (277)
                        |+.++++++.....                                 .-.|.+++|-++-++.           +++++
T Consensus      117 ---L~~~n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~as~~Ag~t~~v~f~l~iPval~~~D~~G~il~qn~~t~l~L~  193 (395)
T P22535         117 ---LHFVNTAKQGAPFLSSMVTDSPIKYGDVMNVIDAPATIAAGATGELTMYYWVPLAYSETDLTGAVLANVPQSKQRLK  193 (395)
T ss_pred            ---HHHHHHHhcCCCeeeeccCCCCcccCCcccccCcceeccCCCeEEEEEEeeeeecCCccceeeeeCCCCccEEEE

Q ss_pred            EEecc---------------eeeecCCCCceeeeceEEEEEEe-----CC-CCeEEE---------eeEccceeeccccC
Q SkuldV1        122 FRFNT---------------IANMNDGDRTTYTGATIDIMVEV-----GT-PGRLSL---------LRNCGKIVLGANTG  171 (277)
Q Consensus      122 i~~~t---------------saq~~gGDr~T~ta~t~DI~v~~~~~~~~kG~~ip-------------g~~s~ga~~G  171 (277)
                     |.+++               .-+..+++..|+..=.+.-+.|+     -. .+...|         +|..-...-.+..-
T Consensus      194 Vt~at~~~~~~~~~~~~a~~y~~~~~t~tl~tv~V~~~~e~~~iPi~~~~~~P~~dlst~~~l~~~~~~~~~G~  273 (395)
T P22535         194 LEFANNNTAFAAVGANPLEAIYQGAGAADCFEEISYTVYQSYLDQLPVGQNGYILPLIDLSTLYNLENSAQAGLTPNVD  273 (395)
T ss_pred            EEEcCccccccccCCChhHHhcCCCcceEEeEEEEEEEEeEeecCCCCCCCccCCCccHHHHHhHhheeeccCCCCce

Q ss_pred            CCcccCCccccceeeeEEEEEecCCCccc--cceeeEE-EeeCceeeeecCcHHHHHHHccCceeeeccceEEEEEecC-
Q SkuldV1        172 RLPEFINPTEFGYKALNLMVLIEDDGTPSN--TAIAKLE-LRKGSNIIREGSIPKLQELTKGKFSIALGTGFVWIPIKQ-  247 (277)
Q Consensus      172 ~l~e~lp~~~~plR~~~Lm~it~D~gT~S~--tay~~L~~Is~g~~~i~DGsiakl~~~~~Ks~vA~s~G~~~i~fp~~  247 (277)
                     +.-++  |...-||.  |++ ++|.|.|.+  +..+.+. +..t...+++-...-+++.++.+.+-.+-.|++++.|-.
T Consensus      274 ~~i~L~~p~g~~yr~~~l~~~v~DnG~~~~~~~d~~~~~l~~n~~di~~~~~~~~~~~~~~~~~~~lp~Gvy~~D~~~  348 (395)
T P22535         274 FVVQY--ANLYRYLS--TIA-VFDNGGSFNAGTDINYLSQRTANFSDTRKLDPKTWAAQTRRRIATDFPKGVYYCDNRDK  348 (395)
T ss_pred            eEEEc--ccccceEE--EEE-EEeCCCcCCCCccccccceeeeCCceeeccCHHHHHHHHHHhCCCCCeEEEEeCCCC

Q ss_pred            ---cCCHHHeEEEEEeccCccceEEEEEee
Q SkuldV1        248 ---AISASQLKLISQIDSAGTNIEVHWMLT  274 (277)
Q Consensus      248 ---kis~~tLkl~~~~tAgt~~e~h~~~~  274 (277)
                        ----+.+.|.+..-+.|.+-.+-+-|.
T Consensus      349 gi~t~~~~~~~l~~~~~~~~~~~~~~~~~  378 (395)
T P22535         349 PIYTLQYGNVGFVVNPKTVNQNARLLMGYE  378 (395)
T ss_pred            ccccccccceEEEEccccccccceeEEEEe

Q ss_pred            HHHHHHHHHHhcccCc-----------------------------cceeEEEcCCccCCCC------------eEEEE
Q SkuldV1         84 IQLMREANKLRNKVADA-----------------------------DEYFEIDFPKAIGRGH------------NAYID  121 (277)
Q Consensus       84 i~lLRE~~~~~~VA~~-----------------------------------d~y~~I~FP~aip~g~------------dvqI~  121 (277)
                        |+.++++++.....                                 .-.|.+++|-++-++.           +++++
T Consensus      117 ---L~~~n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~as~~Ag~t~~v~f~l~iPval~~~D~~G~il~qn~~t~l~L~  193 (395)
T P22535         117 ---LHFVNTAKQGAPFLSSMVTDSPIKYGDVMNVIDAPATIAAGATGELTMYYWVPLAYSETDLTGAVLANVPQSKQRLK  193 (395)
T ss_pred            ---HHHHHHHhcCCCeeeeccCCCCcccCCcccccCcceeccCCCeEEEEEEeeeeecCCccceeeeeCCCCccEEEE

Q ss_pred            EEecc---------------eeeecCCCCceeeeceEEEEEEe-----CC-CCeEEE---------eeEccceeeccccC
Q SkuldV1        122 FRFNT---------------IANMNDGDRTTYTGATIDIMVEV-----GT-PGRLSL---------LRNCGKIVLGANTG  171 (277)
Q Consensus      122 i~~~t---------------saq~~gGDr~T~ta~t~DI~v~~~~~~~~kG~~ip-------------g~~s~ga~~G  171 (277)
                     |.+++               .-+..+++..|+..=.+.-+.|+     -. .+...|         +|..-...-.+..-
T Consensus      194 Vt~at~~~~~~~~~~~~a~~y~~~~~t~tl~tv~V~~~~e~~~iPi~~~~~~P~~dlst~~~l~~~~~~~~~G~  273 (395)
T P22535         194 LEFANNNTAFAAVGANPLEAIYQGAGAADCFEEISYTVYQSYLDQLPVGQNGYILPLIDLSTLYNLENSAQAGLTPNVD  273 (395)
T ss_pred            EEEcCccccccccCCChhHHhcCCCcceEEeEEEEEEEEeEeecCCCCCCCccCCCccHHHHHhHhheeeccCCCCce

Q ss_pred            CCcccCCccccceeeeEEEEEecCCCccc--cceeeEE-EeeCceeeeecCcHHHHHHHccCceeeeccceEEEEEecC-
Q SkuldV1        172 RLPEFINPTEFGYKALNLMVLIEDDGTPSN--TAIAKLE-LRKGSNIIREGSIPKLQELTKGKFSIALGTGFVWIPIKQ-  247 (277)
Q Consensus      172 ~l~e~lp~~~~plR~~~Lm~it~D~gT~S~--tay~~L~~Is~g~~~i~DGsiakl~~~~~Ks~vA~s~G~~~i~fp~~  247 (277)
                     +.-++  |...-||.  |++ ++|.|.|.+  +..+.+. +..t...+++-...-+++.++.+.+-.+-.|++++.|-.
T Consensus      274 ~~i~L~~p~g~~yr~~~l~~~v~DnG~~~~~~~d~~~~~l~~n~~di~~~~~~~~~~~~~~~~~~~lp~Gvy~~D~~~  348 (395)
T P22535         274 FVVQY--ANLYRYLS--TIA-VFDNGGSFNAGTDINYLSQRTANFSDTRKLDPKTWAAQTRRRIATDFPKGVYYCDNRDK  348 (395)
T ss_pred            eEEEc--ccccceEE--EEE-EEeCCCcCCCCccccccceeeeCCceeeccCHHHHHHHHHHhCCCCCeEEEEeCCCC

Q ss_pred            ---cCCHHHeEEEEEeccCccceEEEEEee
Q SkuldV1        248 ---AISASQLKLISQIDSAGTNIEVHWMLT  274 (277)
Q Consensus      248 ---kis~~tLkl~~~~tAgt~~e~h~~~~  274 (277)
                        ----+.+.|.+..-+.|.+-.+-+-|.
T Consensus      349 gi~t~~~~~~~l~~~~~~~~~~~~~~~~~  378 (395)
T P22535         349 PIYTLQYGNVGFVVNPKTVNQNARLLMGYE  378 (395)
T ss_pred            ccccccccceEEEEccccccccceeEEEEe

Q ss_pred            HHHHHHHHHHhcccCc-----------------------------cceeEEEcCCccCCCC------------eEEEE
Q SkuldV1         84 IQLMREANKLRNKVADA-----------------------------DEYFEIDFPKAIGRGH------------NAYID  121 (277)
Q Consensus       84 i~lLRE~~~~~~VA~~-----------------------------------d~y~~I~FP~aip~g~------------dvqI~  121 (277)
                        |+.++++++.....                                 .-.|.+++|-++-++.           +++++
T Consensus      117 ---L~~~n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~as~~Ag~t~~v~f~l~iPval~~~D~~G~il~qn~~t~l~L~  193 (395)
T P22535         117 ---LHFVNTAKQGAPFLSSMVTDSPIKYGDVMNVIDAPATIAAGATGELTMYYWVPLAYSETDLTGAVLANVPQSKQRLK  193 (395)
T ss_pred            ---HHHHHHHhcCCCeeeeccCCCCcccCCcccccCcceeccCCCeEEEEEEeeeeecCCccceeeeeCCCCccEEEE

Q ss_pred            EEecc---------------eeeecCCCCceeeeceEEEEEEe-----CC-CCeEEE---------eeEccceeeccccC
Q SkuldV1        122 FRFNT---------------IANMNDGDRTTYTGATIDIMVEV-----GT-PGRLSL---------LRNCGKIVLGANTG  171 (277)
Q Consensus      122 i~~~t---------------saq~~gGDr~T~ta~t~DI~v~~~~~~~~kG~~ip-------------g~~s~ga~~G  171 (277)
                     |.+++               .-+..+++..|+..=.+.-+.|+     -. .+...|         +|..-...-.+..-
T Consensus      194 Vt~at~~~~~~~~~~~~a~~y~~~~~t~tl~tv~V~~~~e~~~iPi~~~~~~P~~dlst~~~l~~~~~~~~~G~  273 (395)
T P22535         194 LEFANNNTAFAAVGANPLEAIYQGAGAADCFEEISYTVYQSYLDQLPVGQNGYILPLIDLSTLYNLENSAQAGLTPNVD  273 (395)
T ss_pred            EEEcCccccccccCCChhHHhcCCCcceEEeEEEEEEEEeEeecCCCCCCCccCCCccHHHHHhHhheeeccCCCCce

Q ss_pred            CCcccCCccccceeeeEEEEEecCCCccc--cceeeEE-EeeCceeeeecCcHHHHHHHccCceeeeccceEEEEEecC-
Q SkuldV1        172 RLPEFINPTEFGYKALNLMVLIEDDGTPSN--TAIAKLE-LRKGSNIIREGSIPKLQELTKGKFSIALGTGFVWIPIKQ-  247 (277)
Q Consensus      172 ~l~e~lp~~~~plR~~~Lm~it~D~gT~S~--tay~~L~~Is~g~~~i~DGsiakl~~~~~Ks~vA~s~G~~~i~fp~~  247 (277)
                     +.-++  |...-||.  |++ ++|.|.|.+  +..+.+. +..t...+++-...-+++.++.+.+-.+-.|++++.|-.
T Consensus      274 ~~i~L~~p~g~~yr~~~l~~~v~DnG~~~~~~~d~~~~~l~~n~~di~~~~~~~~~~~~~~~~~~~lp~Gvy~~D~~~  348 (395)
T P22535         274 FVVQY--ANLYRYLS--TIA-VFDNGGSFNAGTDINYLSQRTANFSDTRKLDPKTWAAQTRRRIATDFPKGVYYCDNRDK  348 (395)
T ss_pred            eEEEc--ccccceEE--EEE-EEeCCCcCCCCccccccceeeeCCceeeccCHHHHHHHHHHhCCCCCeEEEEeCCCC

Q ss_pred            ---cCCHHHeEEEEEeccCccceEEEEEee
Q SkuldV1        248 ---AISASQLKLISQIDSAGTNIEVHWMLT  274 (277)
Q Consensus      248 ---kis~~tLkl~~~~tAgt~~e~h~~~~  274 (277)
                        ----+.+.|.+..-+.|.+-.+-+-|.
T Consensus      349 gi~t~~~~~~~l~~~~~~~~~~~~~~~~~  378 (395)
T P22535         349 PIYTLQYGNVGFVVNPKTVNQNARLLMGYE  378 (395)
T ss_pred            ccccccccceEEEEccccccccceeEEEEe
```

Figure S3. HHsearch profile-profile comparisons between the putative major capsid protein of skuldvirus SkuldV1 and the double jelly-roll major capsid proteins of (a) corticovirus PM2, (b) turrivirus STIV, and (c) tectivirus PRD1. H(h), α-helix; E(e), β-strand; C(c), coil.

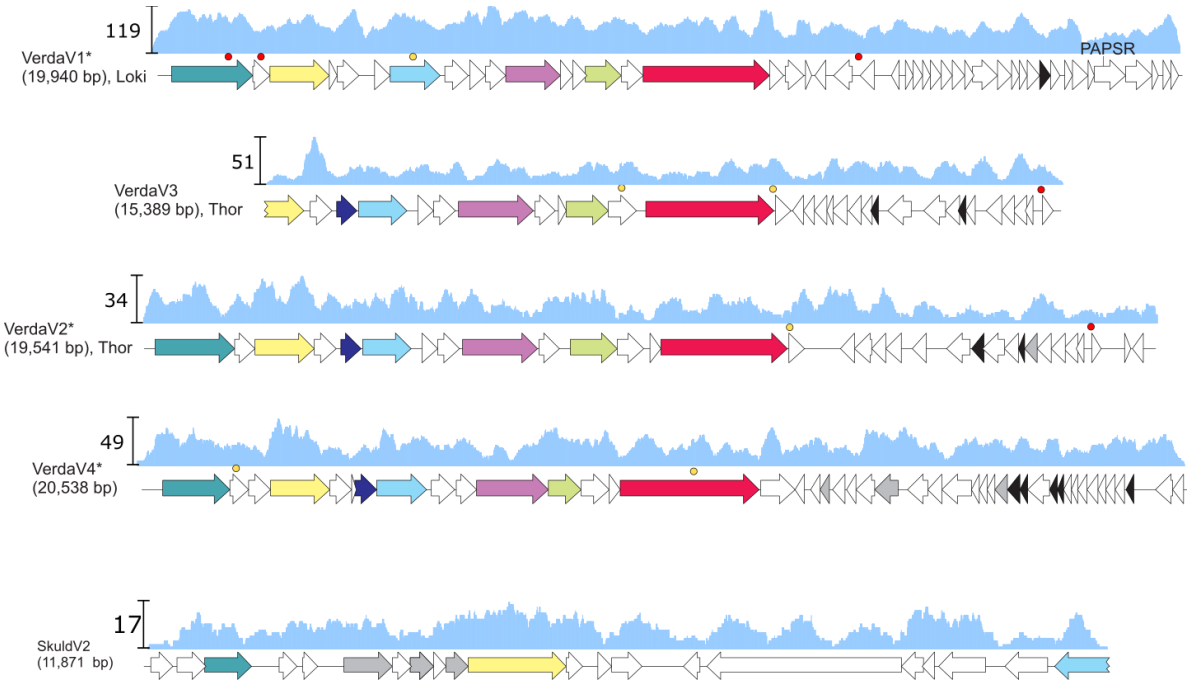


Figure S4. Read depth along viral contigs. The scale on the left shows the maximal coverage for each contig. The open reading frames are colored the same way as in the corresponding Figures 3 and 4 for verdandiviruses and skuldviruses, respectively. Colored circles represent the positions of protospacers targeted by asgardarchaeal CRISPR spacers.

Supplementary Data 1. Output of CRISPRDetect output for CRISPR arrays analyzed in this work.